104IB26

by Cde Anu

Submission date: 26-Nov-2025 10:51AM (UTC+0530)
Submission ID: 2828042920
File name: 104IB26.pdf (17.96M)
Word count: 71725
Character count: 329937

QUANTITATIVE TECHNIQUES & OPERATIONS RESEARCH

M.B.A (IB) First Year

Semester – I, Paper-IV

Director, I/c

Prof. V. VENKATESWARLU

MA.,M.P.S., M.S.W.,M.Phil., Ph.D.

CENTREFORDISTANCEEDUCATION

ACHARAYANAGARJUNAUNIVERSITY

NAGARJUNANAGAR – 522510

Ph:0863-2346222,2346208,

0863-2346259(Study Material)

Website:www.anucde.info

e-mail:anucdedirector@gmail.com

M.B.A (IB) – Quantitative Techniques & Operations Research

First Edition2025

No. of Copies :

This book is exclusively prepared for the use of students of M.B.A. (IB) Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited Circulation only.

Published by:

Prof. V. VENKATESWARLU,

Director, I/C

Centre for Distance Education, Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A^{+}' grade from the NAAC in the year2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the doorstep of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof .K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University

104IB26: Quantitative Techniques & Operations Research

Course Objectives:

1: To expose the students to various statistical and Operations research tools for data analysis.
2. To enable the students to interpretation the results.
3. To facilitate them to take objective decisions based on the models
4 To enable the students to statistical tests for better decision making.
5. To introduce Operations research tools and optimization techniques

Program Outcomes:

1. Gains knowledge of various statistical and Operations Research tools for data analysis.
2. It helps to interpret the results and to take objective decision based on the models.
3. Gains practical knowledge about sampling and sampling methods
4. Better understanding of how to do Analysis
5. It helps understanding and solving real-time Queuing Problems.

UNIT I. Measures of Central Tendency: Arithmetic Mean, Weighted Arithmetic Mean, Median, Mode; Measurement of Variance: Range, Quartile deviation, Average deviation, Standard deviation, Coefficient of variance; Probability: Concept and theorems, Binomial, Poisson and Normal distribution; Central limit theorem.

UNIT II. Correlation, Pearson Correlation, Spearman's Rank Correlation, Regression and Applications, Time Series Analysis.

UNIT III. Hypotheses testing: Errors in testing, one tail & two tail testing, Chi Square test, one sample t-test and two sample t tests, paired t test, Z test, F test ;ANOVA : one way and two way.

UNIT-IV: Operations Research: Linear Programming Basic Concepts, Linear Programming Problem. Graphical Solutions of LPP: Simplex Method - Transportation Problems, Assignment problems.

UNIT-V: Waiting Line Models: Single Channel, Poisson arrival and exponential service times, M/M/l single server system- Economic analysis of waiting line systems-Problems-applications to Inventory Models- Game theory: Terminologies-Two persons zero sum game-Dominance property 2 x n and n x 2 games.

Books Recommended:

1. Gupta S.C: Fundamentals of Business statistics, Sultan Chand, New Delhi.
2. Sancheti and Kapoor V.K., Business Mathematics Sultan Chand & Sons, New Delhi.
3. Hamdy A Taha.A. – Operations Research (Macmillan Publishing)
4. J.K. Sharma – Operations Research (Trinity Pearson).
5. Vijaya Vani Pachala : Fundamentals of Quantitative Techniques, IMRF International Publications, India

CONTENTS

LESSON-1

PROBABILITY THEORY

OBJECTIVES:

After studying this unit, you should be able to:

- Define and give examples of random experiment and trial;
- Define and give examples of sample space, sample point and event;
- Explain mutually exclusive, equally likely, exhaustive and favourable cases and why they are different in nature and how much these terms are important to define probability;
- Explain the classical definition of probability;
- Solve simple problems based on the classical definition of probability;.

STRUCTURE:

1.1 .INTRODUCTION TO PROBABILITY THEORY:

A probability is a numerical value that indicates the likelihood of a specific event occurring, such as experiencing a rainy day, receiving a defective product, or the fluctuation in stock prices. Ya-Lin Chou defines probability as "the science of decision making with calculated risks in the face of uncertainty." When the probability of success is determined based on prior knowledge of the involved processes, it is referred to as a priori probability. In contrast, empirical probability relies on observed data. Subjective probability varies from individual to individual. Below are the fundamental terms utilized in probability theory:

1.Random Experiment: An experiment is classified as a random experiment if, under identical conditions, the outcome is not fixed and can be any one of the possible results. Examples include tossing a coin, rolling a die, or selecting a card from a deck

2. Outcome: The result obtained from a random experiment is termed an outcome.

3. Trial and Event: A single execution of a random experiment is known as a trial, while the outcomes or combinations of outcomes are referred to as events. For instance, when a coin is tossed multiple times, the possible results are heads or tails. Here, tossing coin is the trial, and obtaining heads or tails constitutes the event. Similarly, rolling a die is a trial, and the outcomes of 1, 2, 3, 4, 5, 6 are the events

4. Exhaustive Events: The complete set of possible outcomes from a random experiment is termed exhaustive events. For example, when tossing a coin, the exhaustive outcomes are heads and tails. In the case of rolling a die, the exhaustive outcomes are 1, 2, 3, 4, 5, and 6.

5. Favorable Events: The number of outcomes that are favorable to a specific event during a trial represents the cases that lead to the occurrence of that event. For instance, when drawing a card from a deck, there are four favorable outcomes for drawing an Ace.

6. Mutually Exclusive Events: Events are considered mutually exclusive if the occurrence of one event prevents the occurrence of any other event, meaning that no two or more of these events can happen simultaneously.

1.2. TYPES OF PROBABILITY:

1.2.1. Mathematical or Classical or Apriori Probability

If a random experiment or a trial results in 'n' exhaustive, mutually exclusive and equally likely outcomes, out of which 'm' are favourable to the occurrence of an event E, then the probability 'p' of occurrence of E, denoted by P(E) is given by

$$p = P(E) = \frac{\text{Number of favourable cases}}{\text{Total number of exhaustive cases}} = \frac{m}{n}$$

Remark:

1. Since $m \geq 0$, $n > 0$ and $m \leq n$, $P(E) \geq 0$ and $P(E) \leq 1$. $\Rightarrow 0 \leq p \leq 1$
2. The non-happening of the event E is called the complementary event of E and is denoted by $\bar{E}$ or E^c. The number of favourable cases of $\bar{E}$ is (n-m). Then, the probability q that E will not happen is given by ,

$$q = P(\bar{E}) = 1-p$$

3. p+q=1

Solved Examples

1. Tossing a single coin – $S = \{H, T\}$; $n(S) = 2$ {$n(S)$ is total number of sample points in S.}
2. Tossing of two coins – $S = \{HH, HT, TH, TT\}$; $n(S) = 4$.
3. Throwing a die - $S = \{ 1, 2, 3, 4, 5, 6 \}$ $n(S) = 6$

1.2.2. Statistical or Empirical Probability

If an experiment is performed repeatedly under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event occurs to the number of trials, as the number of trial becomes indefinitely large, is called the probability of happening of the event,

$$P(E) = \lim_{N \to \infty} \frac{M}{N}$$

Sample space: The set of all possible outcomes of a given random experiment is called the sample space associated with that experiment. Each possible outcome or element in a sample space (S) is called a sample point or an elementary event.

Event: Every non-empty subset of S, which is a disjoint union of single element subsets of the sample space S of a random experiment E is called an event.

Problem: Out of 20 employees in a company, 5 are post graduates. Three employees are

selected at random. Find the probability that all the three are post graduates.

$P(E) = {}^{5}C_3/{}^{20}C_3$

Problem: A coin is tossed two times. The sample space is S = {HH, HT, TH, TT}, where H and T denote head and tail. What is the probability of getting at least one head?

Let A be an event of getting at least one head. A = {HH, HT, TH}
P(A) = P(HH) + P(HT) + P(TH)
= 1/4 + 1/4 + 1/4

P(A) = ¾

Problem: If two television tubes are picked in succession from a shipment of 240 television tubes of which 15 tubes are defective. What is the probability that they will both be defective?
Prob. (both the selected tubes will be defective) = 15/240 x 14/239 = 7/1912

Problem: From 25 tickets, marked with the first 25 numerals, one is drawn at random. Find the probability that
i)it is a multiple of 5 or of 7, ii)it is a multiple of 3 or of 7.

i) Numbers (out of the first 25 numerals) which are **multiples of 5** are **5, 10, 15, 20** and **25**, and the numbers which are **multiples of 7** are **7, 14 and 21.**
Pr (a multiple of 5 or 7) = 8 / 25
ii) Numbers (among the first 25 numerals) which are **multiples of 3** are **3, 6, 9, 12, 15, 18, 21, 24** and the numbers which are **multiples of 7** are **7, 14, 21**.
Pr (a multiple of 3 or 7) = 10 / 25 = 2 / 5.

Problem: If two dice are thrown, what is the probability that the sum is

a) Greater than 8 *b)* neither 7 or 11

a) Let *S* denote the sum of the two dice.

$$P(S > 8) = P(S = 9) + P(S = 10) + P(S = 11) + P(S = 12)$$

$$P(S > 8) = \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36} = \frac{5}{18}$$

b) Let *A* denote the event of getting the sum of 7 and *B* denote the event of getting the sum of 11.

$$P(A) = \frac{1}{6}, \quad P(B) = \frac{1}{18}$$

$$P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B) = 1 - P(A) - P(B)$$

$$= 1 - \frac{1}{6} - \frac{1}{18} = \frac{7}{9}$$

Problem: A factory discovers that, on average, 20% of the bolts produced by a given machine are defective for specific specifications. If ten bolts are chosen at random from the day's production of this machine. Determine the likelihood that

a) exactly two items will be faulty.
b) Two or more items will be faulty.
c) More than one will be defective.

Let X be number of defective bolts.

a) $P(X = 2) = {}^{10}C_2(0.2)^2(0.8)^8 = 45(0.04)(0.1678) = 0.3020$

b) $P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$

$= 1 - {}^{10}C_0(0.2)^0(0.8)^{10} - {}^{10}C_1(0.2)^1(0.8)^0$

$= 1 - (0.8)^{10} - 10(0.2)(0.8)^9$

$= 1 - 0.1074 - 0.2684 = 0.6242$

c) $P(X > 5) = P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$

$= {}^{10}C_6(0.2)^6(0.8)^4 + {}^{10}C_7(0.2)^7(0.8)^3 + {}^{10}C_8(0.2)^8(0.8)^2 + {}^{10}C_9(0.2)^9(0.8)^1 + {}^{10}C_0(0.2)^{10}$

$= 0.00637$

Results on probability:

1. $P(\emptyset) = 0$.
2. $P(\bar{A}) = 1 - P(A)$.
3. $P(\bar{A} \cap B) = P(B) - P(A \cap B)$.
4. $P(A \cap \bar{B}) = P(A) - P(A \cap B)$.

+

1.4 ADDITION THEOREM OF PROBABILITY:

If A and B are any two events (subsets of sample space S) and are not disjoint, then

$$P(A \cap B) = P(A) + P(B) - P(A \cap B).$$

Proof:

$$P(A \cup B) = \frac{n(A \cup B)}{n(s)} = n(A) + n(B) - n(A \cap B)$$

$$= \frac{n(A)}{n(s)} + \frac{n(B)}{n(s)} - \frac{n(A \cap B)}{n(s)} = P(A) + P(B) - P(A \cap B)$$

*Corollary*1: If the events A and B are mutually disjoint then

$$A \cap B = \emptyset \quad \Rightarrow \quad P(A \cap B) = P(\emptyset) = 0$$

*Corollary*2: For three non-mutually exclusive events to A, B and C

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Problem: The chance of winning the race by person A is 1/5 and that of person B is 1/6. What is the probability that the race will be won by A or B?

$P(A \cup B) = P(A) + P(B) = 11/30$

Problem:: Find the probability of a 4 turning up at least once in two tosses of a die.

Let A_1 be an event of '4' turning up on the first toss.
Let A_2 be an event of '4' turning up on the second toss.
Since, A_1 and A_2 are not mutually exclusive,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

$$= \frac{1}{6} + \frac{1}{6} - \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)$$

$$P(A_1 \cup A_2) = \frac{11}{36}$$

Problem: A problem in Mathematics is given to three students X, Y and Z whose chances of solving it are 3/4, 1/2 and 1/4 respectively. What is the probability that the problem will be solved?

$$P(X \cup Y \cup Z) = P(X) + P(Y) + P(Z) - P(X \cap Y) - P(X \cap Z) - P(Y \cap Z) + P(X \cap Y \cap Z)$$

$$= \frac{3}{4} + \frac{1}{2} + \frac{1}{4} - \frac{3}{4} \times \frac{1}{2} - \frac{3}{4} \times \frac{1}{4} - \frac{1}{2} \times \frac{1}{4} + \frac{3}{4} \times \frac{1}{2} \times \frac{1}{4} = \frac{29}{32}$$

Problem: In a city (based on a sample survey), the probabilities that a family owns a television set, a washing machine or both television and washing machine are 0.86, 0.35 and 0.29 respectively. What is probability that a family owns either or both?

Let A be an event that a family owns a television set.
Let B be an event that a family owns a washing machine.
P(A) = 0.86
P(B) = 0.35

P(A ∩ B) = 0.29

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$= 0.86 + 0.35 - 0.29$

$= 0.92$

Problem A can hit a target 3 times in 5 shots, B 2 times in 5 shots, and C 3 times in 4 shots. Find the probability of the target being hit at all when all of them try.

Let E_1 be the event that A hits the target.

$$P(E_1) = \frac{3}{5} \text{ and } P(\bar{E}_1) = 1 - \frac{3}{5} = 2/5$$

and E_2 be the event that B hits the target,

$$P(E_2) = \frac{2}{5} \quad and \quad P(\bar{E}_2) = 1 - \frac{2}{5} = 3/5$$

and E_3 be the event that B hits the target,

$$P(E_3) = \frac{3}{4} \quad and \quad P(\bar{E}_3) = 1 - \frac{3}{4} = \frac{1}{4}$$

The required probability 'p' that the target is hit when they all try is given by

$$p = P[atleast\ one\ of\ the\ three\ hits\ the\ target]$$

$$= 1 - P[none\ hits\ the\ target]$$

$$= 1 - P(\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_3)$$

$$= 1 - P(\bar{E}_1)P(\bar{E}_2)P(\bar{E}_3),$$

by compound probability theorem, since E_1, E_2 and E_3 are independent

$\Rightarrow$ $\bar{E}_1$, $\bar{E}_2$ *and* $\bar{E}_3$ are also independent

p = 1-(2/5*3/5*1/4)=0.94

1.5 CONDITIONAL PROBABILITY:

Let *P(A)* represent the likelihood that a random experiment will result in an outcome in the set A relative to the sample space S of the random experiment. If we have prior information that the outcome of the random experiment must be in a set B of S then this information must be used to re-appraise the likelihood that the outcome will also be in B. This re-appraised probability is denoted by

P(A|B) = conditional probability of the event A given that event B has already happened When we know that a particular event B has occurred instead of S, we concentrate our attention on B only and the conditional probability of A given B will be the ratio of that part of A which is included in B (ie $A \cap B$) to the probability of B.

e.g : Let us consider a random experiment of drawing a card from a pack of cards.

A: drawing a king card ; $n(A) = 4$

$$P(A) = \frac{4}{52} = \frac{1}{13}$$

Now suppose that a card is drawn and we are informed that the drawn card is red. How does this information affect the likelihood of the event A?

B: the card drawn is red ; $n(B) = 26$

Now, the *P(A)* must be computed relative to the new sample space B which consists of 26

sample points (red cards only).
Among the 26 red cards, 2(red) kings so $n(A \cap B) = 2$.

Hence the required probability

$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{2}{26} = \frac{1}{13}$$

1.6. MULTIPLICATION THEOREM OF PROBABILITY:

$$P(A \cap B) = \begin{cases} P(A).\,P(A|B), P(A) > 0 \\ P(B).\,P(A|B), P(B) > 0 \end{cases}$$

where $P(B|A)$ represents conditional probability of occurrence of B when the event A has already happened and $P(A|B)$ is the conditional probability of happening of A, given that B has already happened.

Independent Events: An event A is said to be independent of another event B, if the conditional probability of A given $P(A|B)$ is equal to the unconditional probability of B (ie) if $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

1.7. MULTIPLICATION THEOREM OF PROBABILITY FOR INDEPENDENT EVENTS:

If A and B are two events with positive probabilities $\{P(A) \neq 0, P(B) \neq 0\}$ then A and B are independent if and only if $P(A \cap B) = P(A).\,P(B)$

Pairwise independent events: The events $A_1, A_2, \ldots\ldots A_n$ are said to be pairwise independent if and only if

$$P(A_i \cap A_j) = P(A_i).\,P(A_j); \quad i \neq j = 1,2,3,\ldots,n$$

Problem Given that P(A) = 1/3, P(B) = 1/4, P(A|B) = 1/6. Find P(B|A) and P(B|A^c).

P(B|A) = P(A|B) P(B) / P(A) = 1/8 , P(B|A^c) = 5/16

Problem Three marbles are drawn successively from a bag containing 5 blue marbles, 4 green marbles and 6 black marbles. Find the probability that they drawn in the order black, green and blue if each ball is (a) replaced (b) not replaced

Let A be an event of drawing black marble on the first draw
Let B be an event of drawing green marble on the second draw
Let C be an event of drawing blue marble on the third draw

a) If each marble is replaced, then A, B and C are independent events

$$P(ABC) = P(A)P(B)P(C) = \frac{6}{15} \times \frac{4}{15} \times \frac{5}{15} = \frac{8}{225}$$

b) If each marble is not replaced, then A, B and C are dependent events.

$$P(ABC) = P(A)P\left(\frac{B}{A}\right) P\left(\frac{C}{AB}\right)$$

$$= \frac{6}{15} \times \frac{4}{14} \times \frac{5}{13} = \frac{4}{91}$$

Problem: A consumer research organization has studied the services under warranty provided by the 50 new-car dealers in a certain city. The following table gives the results of the findings of the study.

	Good service under warranty	Poor service under warranty
Dealer in business(5 years or more)	16	4
Dealer in business (less than 5 years)	10	20

What is the probability that a customer (who randomly selects one of the new-car dealers), gets the dealer who provides good service under warranty?

(i) What is the probability a customer (who randomly selects the dealer who has been in the business for 5 years or more) gets the dealer who provides good service under warranty?

(ii) What is the probability that one of the dealers who has been in business less than 5 years will provide good service under warranty?

Let A denote the selection of a dealer who provides good service under warranty.
Let D denote the selection of a dealer who has been in business for 5 years or more.

(i) $P(A) = (16 + 10) / 50 = 0.52$

(ii) $P(A| D) = 16 / 20 = 0.80$

(iii) $P(A \mid D^c) = P(A \cap D^c) / P(D^c) = \frac{\frac{10}{50}}{\frac{30}{50}} = 0.20 / 0.60 = 1/3 = 0.33$

Problem Find the probabilities of getting

(a) Three heads in three tosses of a coin

(b) Four sixes and then another number in five throws of a die.

(a) Pr.(Three heads in three tosses of a coin) = 1/2 x 1/2 x 1/2 = 1/8

(b) Pr. (Four sixes and then another number in five throws of a die)

= 1/6 x 1/6 x 1/6 x 1/6 x 5/6

= 5 / 7776

Problem A coin is tossed three times and the outcomes are HHH, HHT, HTH, HTT, THH, THT, TTH, TTT. If A is the event that a head occurs on each of the first two tosses, B is the event that a tail occurs on the third toss and C is the event that exactly two tails occur in three tosses. Show that

(a) Events A and B are independent

(b) Events B and C are dependent

A = {HHH, HHT}
P(A) = 1/4

B = {HHT, HTT, THT, TTT}
P(B) = 1/2
C = {HTT, THT, TTH}
P(C) = 3/8

$A \cap B = \{HHT\}$
$P(A \cap B) = 1/8$
$B \cap C = \{HTT, THT\}$

$P(B \cap C) = 1/4$

(a) Since $P(A) \cdot P(B) = 1/4 \cdot 1/2 = 1/8 = P(A \cap B)$,

Events A and B are independent.

(b) Since $P(B) \cdot P(C) = 1/2 \cdot 3/8 = 3/16 \neq P(B \cap C)$,

Events A and B are dependent.

1.8. BAYES' THEOREM:

If $E_1, E_2, \ldots \ldots E_n$ are mutually disjoint event with $P(E_i) \neq 0$ $(i=1,2,\ldots..,n)$, then for any arbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have

$$P(E_i|A) = \frac{P(E_i)P(A|E_i)}{\sum_{i=1}^{n} P(E_i)P(A|E_i)} = \frac{P(E_i)P(A|E_i)}{P(A)}, \; i = 1,2, \ldots, n$$

Remark: The probabilities $P(E_1). P(E_2) \ldots \ldots P(E_n)$ are termed as 'a prior' probabilities.

1. $P(E_i|A)$ are called 'likelihood'.
2. $P(E_i|A)$ are called 'posterior' probabilities.
3. If the events constitutes a disjoint partition of the sample space S and $P(E_i) \neq 0$; $i=1,2,\ldots,n$ then for any event A in S, we have

$$P(A) = \sum_{i=1}^{n} P(E_i)P(A|E_i)$$

Problem: Basket I contains 1 white, 2 black and 3 red balls. Basket II contains 2 white, 1 black and 1 red ball. Basket III contains 4 white, 5 black and 3 red balls. One basket is chosen at random and two balls are drawn. They happen to be white and red. What is the probability that the balls are drawn from Basket I, II or III?

Let E_1, E_2, E_3 be the event of choosing Basket I, II and III.

Let A be an event that the two balls drawn from the selected basket are white and red.

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$$

$$P(A|E_1) = \frac{1 \times 3}{6C_2} = \frac{1}{5}, \quad P(A|E_2) = \frac{2 \times 1}{4C_2} = \frac{1}{3}, \quad P(A|E_3) = \frac{4 \times 3}{12C_2} = \frac{2}{11}$$

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)} = \frac{33}{118}$$

$$P(E_2|A) = \frac{P(E_2)P(A|E_2)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)} = \frac{55}{118}$$

$$P(E_3|A) = \frac{P(E_3)P(A|E_3)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)} = \frac{30}{118}$$

Problem: An industrial unit has 3 machines- I, II and III which produce the same item. Machines I and II, each produce 30% of the total output, Machine III produces 40% of the remaining output. 2% of Machines I defective while Machine II and III, each produces 3% defective items. All the items are put into one stockpile, and then one item is chosen at random. Find the probability that

a) the selected item is defective.

b) What is the probability that the selected defective item was produced by Machine I?

Let E_1, E_2, E_3 be the events that the item chosen is produced by Machines I, II and III respectively. Let A be an event that the item chosen is defective.

a) $P(A) = P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)$

$= (0.3) \times (0.02) + (0.03) \times (0.03) + (0.4) \times (0.03)$

$= 0.027$

b) $P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(A)} = \frac{(0.02)\times(0.3)}{0.027}$

$= 0.223$

Problem: The members of a consulting firm rent cars from three rental agencies: 60 percent from agency 1, 30 percent from agency 2 and 10 percent from agency 3. If 9 percent of the cars from agency one need a tune-up , 20 percent of the cars from agency 2 need a tune-up and 6 percent of the cars from agency 3 need a tune-up.

(a) What is the probability that a rental car delivered to the firm will need a tune-up?

(b) If a rental car delivered to the consulting firm needs a tune-up, what is the probability that it came from agency 2?

Let E_1, E_2, E_3 be the events that the car comes from rental agencies 1, 2 and 3 respectively. Let A be an event that the car needs a tune-up.

a) $P(A) = P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)$

$= (0.6) \times (0.09) + (0.3) \times (0.20) + (0.1) \times (0.06)$

$= 0.12$

b) $P(E_2|A) = \frac{P(E_2)P(A|E_2)}{P(A)} = \frac{(0.3)\times(0.2)}{0.12} = 0.5$

1.9. RANDOM VARIABLE:

A random variable is a function X(ω) with domain S and range (-∞, ∞) such that for every real number *a*, the event [ω : X(ω) ≤ *a*] ∈ $\mathcal{B}$. In other words, we are considering a function whose domain is the set of possible outcomes, and whose range is a subset of the set of real.

Consider a random experiment (*E*) consisting of two tosses of a coin.
Outcomes: *HH HT TH TT*
Let *X* be a real number associated with the outcome of the experiment. Let *X* be number of heads.

X : 2 1 1 0

Random variables are denoted by capital letters *X, Y, Z,*……

Let *X* be a r.v. with image set $X(S) = \{1, 2, 3, 4, 5, 6\}$

$p(X=1) = P\{1,1\} = 1/36$ $P(X=2) = P\{(2,1),(2,2),(1,2)\} = 3/36$

1.10 SUMMARY:

Let us now summarize the main points which have been covered in this unit.

1. An experiment in which all the possible outcomes are known in advance but we cannot predict as to which of them will occur when we perform the experiment is called random experiment. Performing an experiment is called trial.
2. Set of all possible outcomes of a random experiment is known as sample space. Each outcome of an experiment is visualised as a sample point and set of one or more possible outcomes constitutes what is known as event. The total number of elements in the sample space is called the number of exhaustive cases and number of elements in favour of the event is the number of favourable cases for the event.
3. Cases are said to be mutually exclusive if the happening of any one of them prevents the happening of all others in a single experiment and if we do not have any reason to expect one in preference to others, then they are said to be equally likely.
4. Classical Probability of happening of an event is the ratio of number of favourable cases to the number of exhaustive cases, provided they are equally likely, mutually exclusive and finite.
5. Odds in favour of an event are the number of favourable cases: number of cases against

the event, whereas Odds against the event are the number of cases against the event : number of cases favourable to the event (S) = {1 , 2, 3, 4, 5, 6}

1.10 TECHNICAL TERMS:

1. Sample Space: The set of all the possible outcomes to occur in any trial
2. Sample Point: It is one of the possible results .
3. Experiment or Trial: A series of actions where the outcomes are always uncertain.
4. Event: It is a single outcome of an experiment.
5. Outcome: Possible result of a trial/experiment .
6. Complimentary event: The non-happening events. The complement of an event A is the event, not A (or A')
7. Impossible Event :The event cannot happen

1.11. SELF ASSESSMENT QUESTIONS:

In a lottery, one has to choose six numbers at random out of the numbers from 1 to 30. He/ she will get the prize only if all the six chosen numbers matched with the six numbers already decided by the lottery committee. Find the probability of wining the prize.
If two coins are tossed then find the probability of getting. (i) (ii) At least one head head and tail (iii) At most one head
If three dice are thrown, then find the probability of getting (i) (ii) triplet sum 5 (iii) sum at least 17 (iv) prime number on first die and odd prime number on second and third dice.
Find the probability of getting 53 Mondays in a randomly selected leap year.

1.12 SUGGESTED READINGS

The following books may be used for more indepth study on the topics dealt within this unit.

1. Levin, R.I. & Rubin, D.S., 1991, Statistics for Management, PHI, New Delhi.
2. Gupta, S.P. 1999, Elementary Statistical Methods, Sultan Chand & Sons, New Delhi.
3. Bhardwaj, R.S. 2001, Business Statistics, Excel Books, New Delhi.
4. Chandan, J.S. Statistics for Business and Economics, Vikas Publishing House Pvt. Ltd., New Delhi.

Dr. Naga Nirmala Rani

LESSON- 2

BIONOMIAL DISTRIBUTION

OBJECTIVES:

Study of the present unit will enable you to:

- Define the Bernoulli distribution and to establish its properties;
- Define the binomial distribution and establish its properties;
- Identify the situations where these distributions are applied;
- Know as to how binomial distribution is fitted to the given data; and
- Solve various practical problems related to these distributions

STRUCTURE:

2.1 MEANING & DEFINITION:

Binomial Distribution is associated with James Bernoulli, a Swiss Mathematician. Therefore, it is also called Bernoulli distribution. Binomial distribution is the probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure. In other words, it is used to determine the probability of success in experiments on which there are only two mutually exclusive outcomes. Binomial distribution is discrete probability distribution.

Binomial Distribution can be defined as follows: "A random variable r is said to follow Binomial Distribution with parameters n and p if its probability function is:

$P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$

Where, P = probability of success in a single trial q = 1 – p n = number of trials
r = number of success in 'n' trials.

2.2 ASSUMPTION OF BINOMIAL DISTRIBUTION:

(Situations where Binomial Distribution can be applied)

Binomial distribution can be applied when:-

1. The random experiment has two outcomes i.e., success and failure.
2. The probability of success in a single trial remains constant from trial to trial of the experiment.

3. The experiment is repeated for finite number of
4. The trials are independent

2.3 PROPERTIES (FEATURES) OF BINOMIAL DISTRIBUTION:

- It is a discrete probability distribution.
- The shape and location of Binomial distribution changes as 'p' changes for a given 'n'.
- The mode of the Binomial distribution is equal to the value of 'r' which has the largest probability.
- Mean of the Binomial distribution increases as 'n' increases with 'p' remaining constant.
- The mean of Binomial distribution is np.
- The Standard deviation of Binomial distribution is $\sqrt{npq}$
- The variance of Binomial Distribution is npq
- If 'n' is large and if neither 'p' nor 'q' is too close zero, Binomial distribution may be approximated to Normal Distribution.
- If two independent random variables follow Binomial distribution, their sum also follows Binomial distribution.

2.4 SOLVED EXAMPLES:

Problem: Six coins are tossed simultaneously. What is the probability of obtaining 4 heads?

$r = 4$

$n = 6\ p = ½$

$q = 1 - p = 1 - ½ = ½$

$\square\ p(r = 4) = {}^{6}C_4 (½)^4 (½)^{6-4}$

$$= \frac{6!}{(6-4)!4!} x (½)^{4+2}$$

$$= \frac{6!}{2!4!} x (½)^6$$

$$= \frac{6 x 5}{2 x 1} x \frac{1}{64}$$

$$= \frac{30}{128}$$

$$= 0.234$$

Sol: $P(r) = {}^{n}C_r p^r q^{n-r}$

Problem: The probability that Sachin scores a century in a cricket match is 1/3. What is the probability that out of 5 matches, he may score century in:

(1) Exactly 2 matches No match

Sol: Here $p = 1/3$, $n = 5$, $q = 2/3$ $*P(r) = {}^{n}C_r p^r q^{n-r}$

Probability that Sachin scores centuary in exactly 2 matches is:

$P(r=2) = {}^5C_2\, 1/3^2\, 2/3^{5-2}$

$= \frac{5!}{(5-2)!2!} \times \frac{1}{9} \times \frac{8}{27}$

$= \frac{5 \times 4}{2 \times 1} \times \frac{1}{9} \times \frac{8}{27}$

$= \frac{160}{486}$

$= \frac{80}{243}$

$= 0.329$

Probability that Sachin scores century in no match is:

$P(r=0) = {}^5C_0\, 1/3^{\,0}\, 2/3^{5-0}$

$\frac{5!}{(5-0)!\,0!} \times 1 \times 2/3^5$

$= 1 \times 1 \times (2/3)^5$

$= \frac{32}{243}$

$= 0.132$

Problem: Consider families with 4 children each. What percentage of families would you expect to have :

(a) Two boys and two girls
(b) At least one boy
(c) No girls
(d) At the most two girls

(a) P(having a boy) = ½

P (having a girl) = ½

n = 4

P (getting 2 boys & 2 girls) = p (getting 2 boys)

$= p(r=2) = {}^4C_2\, (½)^2\, (1/2)^{4-2}$

$= \frac{4!}{(4-2)!\,2!} \times (1/2)^2 \times (½)^2$

$= \frac{4 \times 3}{2} \times (1/2)^4$

$= 6 \times 1/16 = 6/16 = 3/8$

∴ Percentage of families with 2 boys and 2 girls = (3/8) x 100 = 37.5 % .

(b) Probability of having at least one boy:

= p (having one boy or having 2 boys or having 3 boys or having 4 boys)

= p (having one boy) + p (having 2 boys) + p (having 3 boys)

+ p (having 4 boys)
= p (r=1) + p (r = 2) + p(r = 3) + p (r = 4)
= 4/16 + 6/16 + 4/46 + 1/16 = 15/16
∴ Percentage of families with at least one boy = (15/16) x 100 = 93.75 %

(c) Probability of having no girls = Probability of having 4 boys
P (r = 4) = 4C4 (½) ^{4} (½) $^{4-4}$ = 1 x (½)^{4} = 1/16
∴ Percentage of families with at least one boy = (1/16) x 100 = 6.25 %

(d) Probability of having at the most 2 girls = P(having 2 or 1 or 0 girls)
= P(having 2 boys or 3 boys or 4 boys)
= 11/16.
∴ Percentage of families with at least one boy = (11/16) x 100
= 68.75 %

Problem: For a binomial distribution mean = 4 and variance = 12/9. Find n.

Sol. Mean np = 4..... (1)
Variance npq = 12/9 ..... (2)
Divide (2) by (1):
We get q = 12/9 ÷ 4 = 12/36 = 1/3
∴ p = 1 - 1/3 = 2/3
∴ n x 2/3 = 4, n = 4 x 3/2 = 6 n = 6

2.5 FITTING A BINOMIAL DISTRIBUTION:

Steps:

1. Find the value of n, p and q
2. Substitute the values of n, p and q in the Binomial Distribution function of $^{n}C_{r}$ $p^{r}q^{n-r}$
3. Put r = 0, 1, 2, ...in the function $^{n}C_{r}$ $p^{r}q^{n-r}$
4. Multiply each of such terms by total frequency (N) to obtain the expected frequency.

Problem: Eight coins were tossed together for 256 times. Fit a Binomial Distribution of getting heads. Also find mean and standard deviation.

Sol: p (getting head in a toss) = ½ , n = 8, q = ½ Binomial Distribution function is
$p(r) = {}^{n}C_{r}\, p^{r}q^{n-r}$
Put r = 0, 1, 2, 3.....8, then are get the terms of the Binomial Distribution.

Binomial Distribution		
No. of Heads (x)	P(x)	Expected Frequency = P(x) x 256
0	$8C_0 (1/2)^0 (1/2)\ 8 =$ 1/256	1
1	$8C_1 (1/2)^1 (1/2)\ 7 =$ 8/256	8
2	$8C_2 (1/2)^2 (1/2)\ 6 =$	28

	28/256	
3	$8C_3 (1/2)^3 (1/2)$ 5 = 56/256	56
4	$8C_4 (1/2)^4 (1/2)$ 4 = 70/256	70
5	$8C_5 (1/2)^5 (1/2)^3$ = 56/256	53
6	$8C_6 (1/2)^6 (1/2)^2$ = 28/256	28
*7	$8C_7 (1/2)^7 (1/2)^1$ = 8/256	8
8	$8C_8 (1/2)^8 (1/2)^0$ = 1/256	1
Total		256

Mean = np = 8*1/2 = 4
S.D = $\sqrt{npq}$ = $\sqrt{8*1/2*\ ½}$ = $\sqrt{2}$ = 1.414

2.6 SUMMARY:

The following main points have been covered in this unit:

- A discrete random variable X is said to follow **Bernoulli distribution** with parameter p if its function is given by
  $P(r) = {}^nC_r p^r q^{n-r}$
  Where, P = probability of success in a single trial
  q = 1 – p
  n = number of trials
  r = number of success in 'n' trials.
- The **constants of Binomial distribution** are: Mean= np,Variance= npq,

2.7 SELF ASSESSMENT QUESTIONS:

1. Define Binomial Distribution.
2. What are the important properties of Binomial Distribution?
3. Examine whether the following statement is true:
   " For a Binomial Distribution, mean = 10 and S D = 4"
4. For a Binomial Distribution, mean = 6 and S D = $\sqrt{2}$. Find parameters. Write down all the terms of the distribution.

2.8. SUGGESTED READINGS:

1. **Beri** G.C. (2007), Business Stastics, (2nd ed.) New Delhi, Tata MCgraw Hill.
2. Levin, J. & Fox, J.A. (2006) Elementary Statistics in Social Research (10th ed.) India, Pearson Education.

Dr.Naga Nirmala Rani

LESSON- 3

POISSON DISTRIBUTION

OBJECTIVES:

After studying this unit, you would be able to:

- know the situations where Poisson distribution is applied;
- define and explain Poisson distribution;
- know the conditions under which binomial distribution tends to Poisson distribution;
- compute the mean, variance and other central moments of Poisson distribution;
- obtain recurrence relation for finding probabilities of this distribution; and
- know as to how a Poisson distribution is fitted to the observed data.

STRUCTURE:

3.1 MEANING AND DEFINITION:

There may be practical situations where the probability of success is very small, that is, there may be situations where the event occurs rarely and the number of trials may not be known.

For instance, the number of accidents occurring at a particular spot on a road everyday is a rare event. For such rare events, we cannot apply the binomial distribution. To these situations, we apply Poisson distribution. The concept of Poisson distribution was developed by a French mathematician, Simeon Denis Poisson (1781-1840) in the year 1837.

Poisson distribution is a limiting form of Binomial Distribution. In Binomial distribution, the total number of trials is known previously. But in certain real life situations, it may be impossible to count the total number of times a particular event occurs or does not occur. In such cases Poisson distribution is more suitable. In case of binomial distributions, as discussed in the last unit, we deal with events whose occurrences and non-occurrences are almost equally important. However, there may be events which do not occur as outcomes of a definite number of trials of an experiment but occur rarely at random points of time and for such events our interest lies only in the number of occurrences and not in its non-occurrences.

Examples of such events are:

i) Our interest may lie in how many printing mistakes are there on each page of a book but we are not interested in counting the number of words without any printing mistake.
ii) In production where control of quality is the major concern, it often requires counting the number of defects (and not the non-defects) per item.
iii) One may intend to know the number of accidents during a particular time interval.

Under such situations, binomial distribution cannot be applied as the value of n is not definite and the probability of occurrence is very small.

Poison Distribution is a discrete probability distribution. It was originated by Simeon Denis Poisson.

A random variable "r" said to follow Binomial distribution if its probability function is:

$$P(r) = \frac{e^{-m} \cdot m^{r}}{r!}$$

Where r = random variable (i.e., number of success in '**n**' trials) e = 2.7183
m = mean of Poisson distribution.

3.2 PROPERTIES OF POISSON DISTRIBUTION:

- Poisson distribution is a discrete probability distribution.
- Poisson distribution has a single parameter 'm'. When 'm' is known all the terms can be found out.
- It is a positively skewed distributionMean and Variance of Poisson distribution are equal to 'm'.

In Poisson distribution, the number of success is relatively small.
Standard deviation of Poisson distribution is $\sqrt{m}$.

3.3 PRACTICAL SITUATIONS WHERE POISSON DISTRIBUTION CAN BE USED:

- To count the number of telephone calls arising at a telephone switch board in a unit of time.
- To count the number of customers arising at the super market in a unit of time.
- To count the number of defects in Statistical Quality Control.
- To count the number of bacteria per unit.
- To count the number of defectives in a park of manufactured goods.
- To count the number of persons dying due to heart attack in a year.
- To count the number of accidents taking place in a day on a busy road.

3.4 SOLVED EXAMPLES:

Problem: A fruit seller, from his past experience, knows that 3 of apples in each basket will be defectives. What is the probability that exactly 4 apples will be defective in a given basket?

Sol. $m = 0.03$

$$P(r) = \frac{e^{-m} \cdot m^r}{r!}$$

$\therefore$ P (exactly 4 apples are defective) = $(e^{-3} \cdot 3^4) / 4!$
$= (0.0498 \times 81) / 24$
$= 0.16807$

Problem: It is known from the past experience that in a certain plant, there are on an average four industrial accidents per year. Find the probability that in a given year there will be less than four accidents. Assume Poisson distribution.

Sol:

$$P(r) = \frac{e^{-m} \cdot m^r}{r!}$$

$m = 4$

$\therefore$ P (exactly 4 apples are defective) = P ($r < 4$)
P ($r < 4$) = P ($r = 0$ or 1 or 2 or 3)
= P ($r = 0$) + P ($r = 1$) + P ($r = 2$) + P ($r = 3$) P ($r = 0$) = $(e^{-4} \cdot 4^0) / 0!$ =
$(0.0183 \times 1) / 1 = 0.0183$
P ($r = 1$) = $(e^{-4} \cdot 4^1) / 1!$ = $(0.0183 \times 4) / 1 = 0.0732$
P ($r = 2$) = $(e^{-4} \cdot 4^2) / 2!$ = $(0.0183 \times 16) / 2 = 0.1464$
P ($r = 3$) = $(e^{-4} \cdot 4^3) / 3!$ = $(0.0183 \times 64) / 6 = 0.1952$
$\therefore$ P ($r < 4$) = 0.0183 + 0.0732 + 0.1464 + 0.1952 = 0.4331

Problem : Out of 500 items selected for inspection, 0.2% is found to be defective. Find how many lots will contain exactly no defective if there are 1000 lots.

Sol:

$$P(r) = \frac{e^{-m} \cdot m^r}{r!}$$

$m = 500 \times 0.2\% = 1$
$\therefore$ P ($r = 0$) = $(e^{-1} \, 1^0) / 0!$ = $(0.3679 \times 1) / 1 = 0.3679$
$\therefore$ No. of lots having zero defective = $0.3679 \times 1000 = 368$

Problem : In a certain factory producing optical lenses, there is a small chance of 1/500 for any one lens to be defective. The lenses are supplied in packets of 10. Use P.D to calculate the approximate number of packets containing no defectives, one defective, two defectives and three defective lenses respectively in a consignment of 20,000 packets.

***Sol*:**

$$P(r) = \frac{e^{-m} \cdot m^r}{r!}$$

$m = 10 \times 1/500 = 0.02$

$\therefore$ P (r = 0) = $(e^{-0.02} \times 0.02^{0}) / 0!$ = (0.9802 x 1) /1 = 0.9802
$\therefore$ No. of packets containing no defective lens = 0.9802 x 20000 = 19604
P (r = 1) = $(e^{-0.02} \times 0.02^{1}) / 1!$ = (0.9802 x 0.02) /1 = 0.0196
$\therefore$ No. of packets containing no defective lens = 0.0196 x 20000 = 392
P (r = 2) = $(e^{-0.02} \times 0.02^{2}) / 2!$ = (0.9802 x 0.0004) /2= 0.00019604
$\therefore$ No. of packets containing no defective lens = 0.00019604 x 20000 = 4
P (r = 3) = $(e^{-0.02} \times 0.02^{3}) / 3!$ = (0.9802 x 0.000008) /6 = 0.0000013069
$\therefore$ No. of packets containing no defective lens = 0.0000013069 x 20000 = 0

3.5 FITTING OF POISSON DISTRIBUTION:

Problem : A Systematic sample of 100 pages was taken from a dictionary and the observed frequency distribution of foreign words per page was found to be as follows:

No. of foreign words per page (x) : 0 1 2 3 4 5 6
Frequency (f) 4 8 2 7 1 2 7 4 11 Calculate the expected frequencies using Poisson distribution.

Sol: At first, we have to know the parameter of P.D, which is equal to the mean of the given distribution. So find the mean of the distribution:

Mean = $(\varepsilon fx) / \varepsilon f$

x	0	1	2	3	4	5	6	
f	48	27	12	7	4	1	1	N = εf = 100
fx	0	27	24	21	16	5	6	(εfx) = 99

Mean = 99/100 = 0.99

Calculation of Expected Frequencies		
X	P(x) = $(e^{-m} . m^{x}) / x!$	Expected Frequency = P(x) . N
0	$(e^{-0.99} . 0.99^{0})/0!$ =(0.3716 x1)/1 = 0.3716	0.3716 x 100 = 37.16 = 37
1	$(e^{-0.99} . 0.99^{1})1!$ =(0.3716 x 0.99)/1= 0.3679	0.3716 x 100 = 37.16 = 37
2	$(e^{-0.99} . 0.99^{2})2!$ =(0.3716 x 0.98)/2= 0.1821	0.1821 x 100 = 18.21 = 18
3	$(e^{-0.99} . 0.99^{3})3!$ =(0.3716 x 0.97)/6= 0.0601	0.0601 x 100 = 6.01 = 6
4	$(e^{-0.99} . 0.99^{4})4!$=(0.3716 x 0.96)24= 0.0149	0.0149 x 100 = 1.49 = 2
5	$(e^{-0.99} . 0.99^{5})5!$=(0.3716 x 0.95)120= 0.0029	0.0029 x 100 = 0.29 = 0

6	$(e^{-0.99} . 0.99^6)6!=(0.3716 \times 0.94)720= 0.0005$	0.0005 x 100 = 0.05 = 0
	Total	**100**

3.6 SUMMARY:

1. A random variable X is said to follow **Poisson distribution** if it assumes indefinite number of non-negative integer values and its probability mass function is given by:

$$p(x) = P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{\lfloor x}; & x = 0, 1, 2, 3, \ldots \text{ and } \lambda > 0. \\ 0; & \text{elsewhere} \end{cases}$$

2. For Poisson distribution, **Mean = Variance** $= \mu_3 = \lambda$, $\mu_4 = 3\lambda^2 + \lambda$

3. **Recurrence relation for probabilities of Poisson distribution** is

$$p(x+1) = \frac{\lambda}{x+1} p(x), \; x = 0, 1, 2, 3, \ldots$$

4. **Expected frequencies for a Poisson distribution** are given by

$$f(x) = N.P[X = x] = N. \frac{e^{-\lambda}\lambda^x}{\lfloor x}; x = 0, 1, 2, \ldots$$

3.7 SELF ASSESSMENT QUESTIONS:

1. Define Poisson distribution.
2. What are the important properties of P.D?
3. What are the situations under which P D can be applied?
4. Write down the probability function of P.D. whose mean is 2. What is its variance?
5. A machine is producing 4% defectives. What is the probability of getting at least 4 defectives in a sample of 50 =, using (a) BD and (b) PD?
6. The following table gives the number of days in a 50 day period during which automobile accidents occurred in a certain part of the city. Fit a Poisson distribution to the data:

No. of accidents	0	1	2	3	4
No. of days	19	18	8	4	1

3.8 SUGGESTED READINGS:

The following books may be used for more indepth study on the topics dealt within this unit.

1. Levin, R.I. & Rubin, D.S., 1991, Statistics for Management, PHI, New Delhi.
2. Gupta, S.P. 1999, Elementary Statistical Methods, Sultan Chand & Sons, New Delhi.
3. Bhardwaj, R.S. 2001, Business Statistics, Excel Books, New Delhi. Chandan, J.S. Statistics for Business and Economics, Vikas Publishing House Pvt. Ltd., New Delhi.

Dr. Naga Nirmala Rani

APPENDIX

VALUE OF $E^{-\lambda}$ (FOR COMPUTING POISSON PROBABILITIES

λ	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	0.9900	0.9802	0.9704	0.9608	0.9512	0.9418	0.9324	0.9231	0.9139
0.1	0.9048	0.8958	0.8860	0.8781	0.8694	0.8607	0.8521	0.8437	0.8353	0.8270
0.2	0.7187	0.8106	0.8025	0.7945	0.7866	0.7788	0.7711	0.7634	0.7558	0.7483
0.3	0.7408	0.7334	0.7261	0.7189	0.7118	0.7047	0.6970	0.6907	0.6839	0.6771
0.4	06703	0.6636	0.6570	0.6505	0.6440	0.6376	0.6313	0.6250	0.6188	0.6125
0.5	0.6065	0.6005	0.5945	0.5886	0.5827	0.5770	0.5712	0.5655	0.5599	0.5543
0.6	0.5448	0.5434	0.5379	0.5326	0.5278	0.5220	0.5160	0.5113	0.5066	0.5016
0.7	0.4966	0.4916	0.4868	0.4810	0.4771	0.4724	0.4670	0.4630	0.4584	0.4538
0.8	0.4493	0.4449	0.4404	0.4360	0.4317	0.4274	0.4232	0.4190	0.4148	0.4107
0.9	0.4066	0.4026	0.3985	0.3946	0.3906	0.3867	0.3829	0.3791	0.3753	0.3716
($\square$=1, 2, 3, ...,10)										
λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$	0.3679	0.1353	0.0498	0.0183	0.0070	0.0028	0.0009	0.0004	0.0001	0.00004

LESSON- 4

NORMAL DISTRIBUTION

OBJECTIVES:

After completion of this unit, you will be able to understand:

- Describe meaning of normal distribution;
- Identify the characteristics of the normal distribution;
- Analyse the properties of the normal distribution; and
- Apply the normal distribution.

STRUCTURE:

4.1 MEANING AND DEFINITION:

The normal distribution is a continuous probability distribution. It was first developed by De-Moivre in 1733 as limiting form of binomial distribution. Fundamental importance of normal distribution is that many populations seem to follow approximately a pattern of distribution as described by normal distribution. Numerous phenomena such as the age distribution of any species, height of adult persons, intelligent test scores of students, etc. are considered to be normally distributed.

Definition of Normal Distribution

A continuous random variable, 'X', said to follow Normal Distribution if its probability function is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

4.2 PROPERTIES OF NORMAL DISTRIBUTION (NORMAL CURVE):

- Normal distribution is a continuous distribution.
- Normal curve is symmetrical about the mean.
- Both sides of normal curve coincide exactly.

- Normal curve is a bell shaped curve.
- Mean, Median and Mode coincide at the centre of the curve.
- Quantities are equi-distant from median. $Q_3 - Q_2 = Q_2 - Q_1$
- Normal curve is asymptotic to the base line.
- Total area under a normal curve is 100%.
- The ordinate at the mean divide the whole area under a normal curve into two equal parts. (50% on either side).
- The height of normal curve is at its maximum at the mean.
- The normal curve is unimodel, i.e., it has only one mode.
- Normal curve is mesokurtic.
- No portion of normal curve lies below the x-axis.
- Theoretically, the range of normal curve is $-\infty$ to $+\infty$. But practically the range is $\mu - 3\sigma$ to $\mu + 3\sigma$.
- Area under the normal curve is distributed as follows: (Area property)
  (a) $\mu \pm \sigma$ covers 68.27% area
  (b) $\mu \pm \sigma$ covers 95.45% area
  (c) $\mu \pm \sigma$ covers 99.73% area

4.3 IMPORTANCE OR USES OF NORMAL DISTRIBUTION:

The normal distribution is of central importance because of the following reasons:

1. The discrete probability distributions such as Binomial distribution and Poisson distribution tend to normal distribution as 'n' becomes large.
2. Almost all sampling distributions conform to the normal distribution for large values of 'n'.
3. Many tests of significance are based on the assumption that the parent population from which samples are drawn follows normal distribution.
4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. Normal distribution finds applications in Statistical Quality Control.
6. Many distributions in social and economic data are approximately normal. For example, birth, death, etc. are normally distributed.

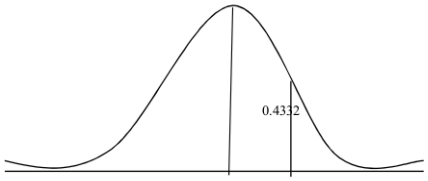
4.4 AREA UNDER STANDARD NORMAL CURVE:

In case of normal distribution, probability is determined on the basis of area. But the area

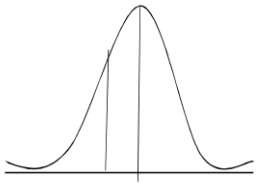
we have to calculate the ordinate of z – scale.
The scale to which the standard deviation is attached is called z-scale.
$Z = (x - \mu) / \sigma$

4.5 SOLVED PROBLEMS:

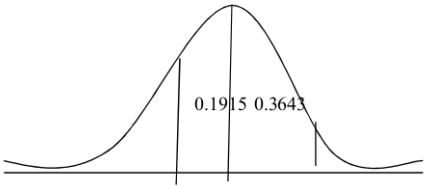
Problem: The variable, x, follows normal distribution with mean = 45 and S.D = 10. Find the probability that $x \geq 60$.
Sol: $\mu = 45$, $\sigma = 10$, $x = 60$
$Z = (x - \mu) / \sigma$ $Z = (60 - 45) / 10 = 15/10 = 1.5$



$P(x \geq 60)$ means $P(z \geq 1.5)$
$= 0.5 - 0.4332 = 0.0668$
$P(x \geq 60) = 0.0668$
Problem: The variable, x, follows normal distribution with mean = 45 and S.D = 10. Find the probability that $x \leq 40$.
Sol: $\mu = 45$, $\sigma = 10$, $x = 40$
$Z = (x - \mu) / \sigma$
$Z = (40 - 45) / 10 = -5/10 = -0.5$
$P(x \leq 40)$ means $P(z \geq -0.5)$
$= 0.5 - 0.1915 = 0.3085$
$P(x \leq 40) = 0.3085$
Problem: The variable, x, follows normal distribution with mean = 45 and S.D = 10. Find the probability that $40 \leq x \leq 56$.
Sol: $\mu = 45$, $\sigma = 10$, $x_1 = 40$, $x_2 = 56$
$Z = (x - \mu) / \sigma$
When $x = 40$, $Z = (40 - 45) / 10 = -5 / 10 = -0.5$



When x = 56, Z = (56 -- 45) / 10 = 11 / 10 = 1.1



P (40 ≤ x ≤ 56) means P (z – 0.5 ≤ x ≤ 1.5)
= 0.1915 + 0.3643 = 0.5558
P (40 ≤ x ≤ 56) = 0.5558

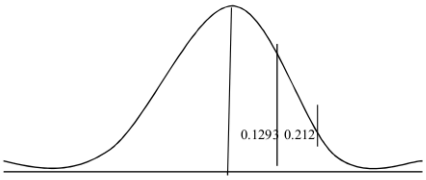
Problem : The scores of students in a test follow normal distribution with mean = 80 and S D = 15. A sample of 1000 students has been drawn from the population. Find (1) probability that a randomly chosen student has score between 85 and 95 (2) appropriate number of students scoring less than 60.

Sol.

(1) $\mu = 80$, $\sigma = 15$, $x_1 = 85$, $x_2 = 95$

(2) $Z = (x - \mu) / \sigma$

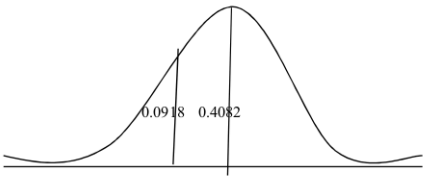
When x = 85, Z = (85 – 80) / 15 = 5/15 = 0.333
When x = 95, Z = (95 – 80) / 15 = 15/15 = 1



∴ P (85 ≤ x ≤ 95) = P (0.333 ≤ z ≤ 1) = 0.3413 – 0.1293 = 0.212
Probability that a student scores between 85 and 95 = 0.212

(3) P (Less than 60):

When x = 60,
Z = (x – μ)/σ = (60 -- 80)/15 = -20/15 = - 1. 333



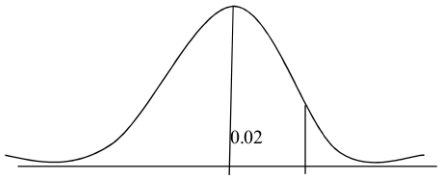
P(x < 60) = P(z < -1.333) = 0.5 -- 0.4082 = 0.0918

∴ Number of students scoring less than 60 = 0.0918 x 1000 = 91.8= 92 students

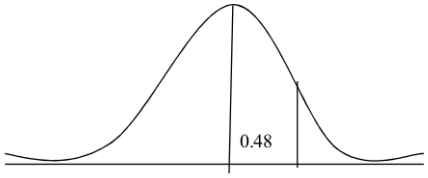
4.6 COMPUTATION OF Z-VALUE WHEN AREA IS KNOWN:

Problem: In a competitive examination, 5000 candidates have appeared. Their average mark was 62 and S.D was 12. If there are only 100 vacancies, find the minimum marks that one should score in order to get selection.

Sol: $= 62, \ \sigma = 12$
Number of vacancies = 100
Percentage of vacancies to the total number of candidates = (100/5000) x 100 = 2% = 0.02
Area corresponds to the students who will get selection is shown in the following normal curve:



Therefore, the area to the left of the above area of 0.02 is: Z = (40 – 45) / 10 = -5/10 = -- 0.5



Locate the area of 0.48 in the table and find the Z – value corresponds to it.
The table shows the area nearest to 0.48 is 0.4798, and the corresponding z-value is 2.05
Z = 2.05
$(x - \mu)/\sigma = 2.05$
$(x - 62)/12 = 2.05, \quad x - 62 = 2.05 \times 12$
$x - 62 = 24.6, \ \therefore \ x = 24.6 + 62 = 86.6$
∴ The minimum marks one should score to get section
= 86.6 marks

4.7 CONSTRUCTION OF NORMAL DISTRIBUTION:

Procedure:

1. Find the mean and S.D of the given distribution and take them as μ and σ (parameters) of the normal distribution.
2. Take the lower limit of each class as the x values.
3. Calculate the z-value corresponding to each x-value by using formulae $z = (x - \mu)/\sigma$. Z-value of first and last values need not be computed.

4. Find the area corresponds to z-value from the standard normal distribution table. The area corresponds to the first and last z-values will be 0.5.
5. Find the area of each class using the area (probability) of respective class limits. (Take the difference in case of same signs; and take the total in case of opposite signs)
6. Multiply the area of each class by the total frequency to the frequency of the class.

The new frequency distribution with theoretical frequencies will be a normal approximation to the given frequency distribution.

Problem : Fit a normal distribution to the following data:

X	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	4	22	48	66	40	16	4

Sol:

Computation of Mean and Standard deviation							
Class	**Mid point (m)**	**F**	**d (m-35)**	**d'**	**fd'**	**d'^{2}**	**fd'^{2}**
10-20	15	4	-20	-2	-8	4	16
20-30	25	22	-10	-1	-22	1	22
30-40	35	48	0	0	0	0	0
40-50	45	66	10	1	66	1	66
50-60	55	40	20	2	80	4	160
60-70	65	16	30	3	48	9	144
70-80	75	4	40	4	16	16	64
		200			180		472

$\bar{x} = A + [(\varepsilon fd')/N] \times C$, $\bar{x} = 35 + [(180/200)x10]$,
$= 35+9 = 44$

S. D $= \sqrt{(\varepsilon fd'^2/N) - [(\varepsilon fd')/N]^2} \times 10 = \sqrt{1.55} \times 10 = 12.45$
$\therefore \mu = 44$ and $\sigma = 12.45$

Computation of Expected Frequencies				
Lower limit	Z= (x—μ)/σ	Area	Area of class	Expected Frequency
10	-2.73	0.5000	0.0268	5
20	-1.93	0.4732	0.1046	21
30	-1.12	0.3686	0.2431	49
40	-0.32	0.1255	0.3099	62
50	0.48	0.1884	0.2171	43
60	1.29	0.4015	0.0802	16
70	2.09	0.4817	0.0183	4
80	2.89	0.5000		
			Total	**200**

4.8 SUMMARY:

In this chapter we introduced the concept of normal distribution. It is a symmetrical bell shaped curve. It is not skewed. Normal curve can be used to determine the percent of the total area under the normal curve associated with any given sigma distance from the mean. Any given sigma distance above the mean contains the identical properties of cases as the same sigma distances below the mean.

4.9 TECHNICAL TERMS:

Normal distribution : A symmerical bell shaped curve. The two tail of the curve never touch the horizontal axis.
Random variable : A variable that assume a unique numerical value for each of the outcomes is a sample space of a probability experiment.
Standard score : Known as Z scores also can be obtained by taking the deviation from mean and divided by standard deviation.

4.10 SELF ASSESSMENT QUESTIONS:

1. Define the normal distribution.
2. What are the key features of normal distribution?
3. Explain the significance of normal distribution.
4. Explain how to generate a normal distribution. 5.
5.Using a normal distribution with a mean of 12 and a variance of 16, calculate the probabili ty of x exceeding 20.
6. 1000 workers' weekly incomes are regularly distributed, with a mean of 70 and a standard deviation of 5. Estimate the number of workers whose pay fall between 69 and 72.
In an aptitude test administered to 900 students, the mean score is 50 and S.D is 20. Find the number of students securing scores (a) between 30 and 70 (b) exceeding 65. Find the value of the score exceeded by the top 90 students.

2. Construct a normal distribution to the following data of marks obtained by 100 students:

Marks	60-62	63-65	66-68	69-71	72-74
No. of Students	5	18	42	27	8

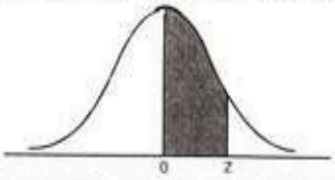
4.11. SUGGESTED READINGS:

- **Beri** G.C. (2007), Business Stastics, (2nd ed.) New Delhi, Tata MCgraw Hill.
- Levin, J. & Fox, J.A. (2006) Elementary Statistics in Social Research (10th ed.) India, Pearson Education.

Dr. Naga Nirmala Rani

VII. AREAS UNDER THE STANDARD NORMAL DISTRIBUTION

The entries in this table are the probabilities that a standard normal variate is between 0 and Z (the shaded area).



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1519	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

LESSON-5

HYPOTHESIS TESTING

OBJECTIVES:

The purpose of studying this lesson is:

- To describe the importance of hypothesis testing in statistical decision-making.
- To formulate null and alternative hypotheses.
- To apply the steps in hypothesis testing to determine whether to accept or reject the null hypothesis.
- To identify and explain Type I and Type II errors, including their implications in hypothesis testing.
- To compute and interpret p-values to support statistical conclusions.
- To interpret key concepts such as confidence level, significance level, and the Power of a test in hypothesis testing outcomes.

STRUCTURE:

5.1 HYPOTHESIS AND HYPOTHESIS TESTING:

A statistical hypothesis is a claim (assertion, statement, belief, or assumption) about an unknown population parameter value. For example: (i) a judge assumes that a person charged with a crime is innocent and subjects this assumption (hypothesis) to verification by reviewing the evidence and hearing testimony before reaching a verdict; (ii) a pharmaceutical company claims the efficacy of a medicine for a disease from which 95 percent of suffering patients are cured; (iii) an investment company asserts that the average return across all its investments is 20 percent, and so on. Sample data are collected and analyzed to test such claims or assertions statistically. Based on sample findings, the hypothesized value of the population parameter is either accepted or rejected. The process that allows a decision-maker to test a claim's validity

(or significance) by analyzing the difference between the value of the sample statistic and the corresponding hypothesized population parameter value is called hypothesis testing.

5.1.1 Formats of Hypothesis

As previously mentioned, a hypothesis is a statement to be tested concerning the actual value of a population parameter using sample statistics. A hypothesis can also be evaluated for any significant differences between two or more populations regarding their standard parameters. A hypothesis can be formulated as an if-then statement to determine whether a difference exists. For example, consider the nature of the following statements:

- If the inflation rate has decreased, the wholesale price index will also decrease.
- If employees are healthy, then they will take sick leave less frequently.

Suppose terms such as 'positive,' 'negative,' 'more than,' 'less than,' etc., are used to make a statement. In that case, this hypothesis is called a directional hypothesis because it indicates the direction of the relationship between two or more populations under study concerning a parameter value, as illustrated below:

- Side effects were experienced by less than 20 percent of people who take a particular medicine.
- The more significant the stress experienced on the job, the lower the employees' job satisfaction.

The nondirectional hypothesis indicates a relationship or difference but does not specify the direction of these relationships or differences. In other words, while it might be clear that a significant relationship exists between two populations for a parameter, we cannot determine whether this relationship is positive or negative. Likewise, even if we recognize that two populations differ in a parameter, it will be challenging to identify which population is greater or lesser. The following examples illustrate nondirectional hypotheses.

- There is a relationship between age and job satisfaction.
- There is a difference between the average pulse rates of men and women.

5.1.2 The Rationale for Hypothesis Testing

Inferential statistics focus on estimating unknown population parameters using sample statistics. When a claim or assumption is made about a specific population parameter value, the corresponding sample statistic is expected to be close to the hypothesized parameter value.

This expectation holds only if the hypothesized parameter value is accurate and the sample statistic accurately estimates the parameter. This method of testing a hypothesis is known as a test statistic.

Since sample statistics are random variables, their sampling distributions exhibit variability. Therefore, we do not expect the value of the sample statistic to equal the hypothesized parameter value. Any difference arises from chance or sampling error. However, if the value of the sample statistic significantly deviates from the hypothesized parameter value, it raises questions about the accuracy of that hypothesized parameter value. The more significant the difference between the sample statistic and the hypothesized parameter, the more doubt there is regarding the validity of the hypothesis.

In statistical analysis, the difference between the value of the sample statistic and the hypothesized parameter is specified in terms of the given probability level, indicating whether the specific level of difference is significant when the hypothesized parameter value is correct.

The likelihood of a particular level of deviation occurring by chance can be calculated from the known sampling distribution of the test statistic.

The probability level at which the decision-maker concludes that the observed difference between the value of the test statistic and the hypothesized parameter value cannot be due to chance is called the *level of significance* of the test.

5.2 GENERAL PROCEDURE FOR HYPOTHESIS TESTING:

As mentioned earlier, to evaluate the validity of the claim or assumption regarding the population parameter, a sample is drawn from the population and analyzed. The analysis results are used to determine whether the claim is valid. The general Procedure for hypothesis testing includes the following summarized steps:

Step 1: State the Null Hypothesis (H_0) and Alternative Hypothesis (H_1)
The null hypothesis H0 represents the claim or statement about a population parameter's value or range of values. The capital letter H stands for hypothesis, and the subscript 'zero' implies 'no difference' between the sample statistic and the parameter value. Thus, hypothesis testing requires that the null hypothesis be considered valid (the status quo or no difference) until it is proven false based on the results observed from the sample data. The null hypothesis is always expressed as a mathematical statement that includes the signs ($\leq$, $=$, $\geq$), claiming the specific value of the population parameter. That is:

Null hypothesis: The hypothesis is initially assumed to be true, although it may be true or false based on the sample data.
H_0: μ ($\leq$, $=$, $\geq$) μ_0
where μ is the population mean and μ_0 represents a hypothesized value of μ. Only one sign out of $\leq$, $=$, and $\geq$ will appear when stating the null hypothesis.

An **alternative hypothesis**, H_1, is the counterclaim (statement) made against the value of the particular population parameter. An alternative hypothesis must be valid when the null hypothesis is false. In other words, the alternative hypothesis states that a specific population parameter value is not equal to the value stated in the null hypothesis and is written as:

Alternative hypothesis: The hypothesis is concluded to be true if the null hypothesis is rejected.

H_1: $\mu \neq \mu_0$
Consequently, H_1: $\mu < \mu_0$ or H_1: $\mu > \mu_0$
Each of the following statements is an example of a null hypothesis and an alternative hypothesis:

- H_0: $\mu = \mu_0$; H_1 : $\mu \neq \mu_0$
- H_0: $\mu \leq \mu_0$; H_1 : $\mu > \mu_0$
- H_0: $\mu \geq \mu_0$; H_1 : $\mu < \mu_0$

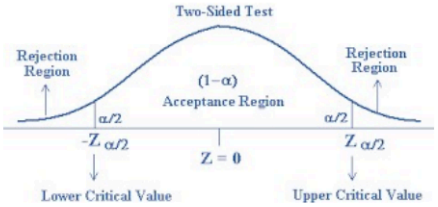
Step 2: State the Level of Significance, α (alpha)
The level of significance, usually denoted by α (alpha), is specified before the samples are drawn, so the results obtained should not influence the decision-maker's choice. The probability of the null hypothesis H0 being wrong is determined. In other words, the significance level defines the likelihood of rejecting a null hypothesis when it is true, i.e., it is the risk a decision-maker takes of rejecting the null hypothesis when it is *true*. The guide provided by the statistical

theory is that this probability must be 'small.' Traditionally, $\alpha = 0.05$ is selected for consumer research projects, $\alpha = 0.01$ for quality assurance, and $\alpha = 0.10$ for political polling.

Step 3: Establish Critical or Rejection Region

The area under the test statistic's sampling distribution curve is divided into two mutually exclusive regions(areas), as shown in Figure 5.1. These regions are called the *acceptance and rejection (or critical) regions*.

Figure 5.1 Areas of Acceptance and Rejection of H_0 (Two-Tailed Test)



The acceptance region is a *range of values* of the sample statistic spread around the *null hypothesized population parameter.* If the values of the sample statistic fall within the limits of the acceptance region, the null hypothesis is accepted. Otherwise, it is rejected.

The **rejection region** is the *range of sample statistic values* within which, if values of the sample statistic fall (i.e., outside the acceptance region's limits), the null hypothesis is rejected.

Rejection region: The range of values leading to rejecting a null hypothesis.

The value of the sample statistic that separates the regions of acceptance and rejection is called the **critical value.**

Critical value: A table value with which a test statistic is compared to determine whether a null hypothesis should be rejected.

The size of the rejection region is directly related to the level of precision with which decisions are made about a population parameter. Decision rules concerning the null hypothesis are as follows:

- If prob (H_0 is true) $\leq \alpha$, then reject H_0
- If prob (H_0 is true) $\geq \alpha$, then accept H_0

In other words, if the probability of H0 being actual is less than or equal to the significance level, α, then reject H0; otherwise, accept H0. The level of significance α is used as the cutoff point separating *the area of acceptance from the area of rejection.*

Step 4: Select a Suitable Test of Significance or Test Statistic

The tests of significance, or test statistics, are classified into parametric and nonparametric tests. Parametric tests are more potent because their data is derived from interval and ratio measurements. Nonparametric tests are utilized to test hypotheses involving nominal and ordinal data. Parametric techniques are preferred, provided certain assumptions are met. The assumptions for parametric tests are as follows:

1. The selection of any element (or member) from the population should not affect the chance for any other to be included in the sample to be drawn from the population.
2. The samples should be drawn from normally distributed populations.
3. Populations under study should have equal variances.

Nonparametric tests have few assumptions and do not specify normally distributed populations or homogeneity of variance.

Selection of a test. For choosing a particular test of significance, the following three factors are considered:

1. Does the test involve one, two, or *k* samples?
2. Are two or more samples used independently or related?
3. Is the measurement scale nominal, ordinal, interval, or ratio?

Further, it is also essential to know (i) sample size, (ii) the number of samples and their size, and (iii) whether data have been weighted. Such questions help in selecting an appropriate test statistic.

One-sample tests are used for a single sample and to test the hypothesis that it comes from a specified population. The following questions need to be answered before using one-sample tests:

- Is there a difference between observed frequencies and the expected frequencies based on a statistical theory?
- Is there a difference between observed and expected proportions?
- Is it reasonable to conclude that a sample is drawn from a population with some specified distribution (normal, Poisson, and so on)?
- Is there a significant difference between some measures of central tendency and their population parameter?

The value of the test statistic is calculated from the distribution of the sample statistic by using the following formula.

$$\text{Test statistic} = \frac{Value\ of\ sample\ statistic - Value\ of\ hypothesized\ population\ parameter}{Standard\ error\ of\ the\ sample\ statistic}$$

The choice of a probability distribution of a sample statistic is guided by the sample size *n* and the value of population standard deviation σ.

Step 5: Formulate a Decision Rule to Accept the Null Hypothesis

Compare the calculated value of the test statistic with the critical value (also called *the standard table value of the* test statistic). The decision rules for the null hypothesis are as follows:

- Accept H_0 if the test statistic value falls within the area of acceptance.
- Reject otherwise

In other words, if the calculated absolute value of a test statistic is less than or equal to its critical (or table) value, then accept the null hypothesis; otherwise, reject it.

5.3 DIRECTION OF THE HYPOTHESIS TEST:

The location of the rejection region (or area) under the sampling distribution curve determines the direction of the hypothesis test, i.e., either lower-tailed or upper-tailed of the sampling

distribution of the relevant sample statistic being tested. It indicates the range of sample statistic values that would lead to a rejection of the null hypothesis. Figures 5.1, 5.2(a), and 5.2(b) illustrate the acceptance region and rejection region of a null hypothesized population mean, μ value, for three different ways of formulating the null hypothesis.

The null hypothesis and the alternative hypothesis are stated as

$H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$

Imply that the sample statistic values, which are either significantly smaller than or greater than the null hypothesized population mean, μ0 value, will lead to rejection of the null hypothesis. Hence, it is necessary to keep the rejection region at both tails of the sampling distribution of the sample statistic. This type of test is called *a two-tailed test or a nondirectional test, as shown in Fig. 5.1. If the significance level for the test is α percent, then the* rejection region equal to α/2 percent is kept in each tail of the sampling distribution.

1. The null hypothesis and alternative hypothesis are stated as

$H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ (Right-tailed test)

or $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$ (Left-tailed test)

Imply that the value of the sample statistic is either 'higher than (or above)' or 'lower than (or below)' than the hypothesized population mean, μ_0 value. This leads to the rejection of the null hypothesis for significant deviation from the specified value μ_0 in one direction (or tail) of the sampling distribution. Thus, the entire rejection region corresponding to the significance level, α percent, lies only in one tail of the sampling distribution of the sample statistic, as shown in Fig. 5.2(a) and 5.2(b). This type of test is called **a one-tailed test or a** *directional test.*

One-tailed test: The test of a null hypothesis that can only be rejected when the sample test statistic value is at one end of the sampling distribution.

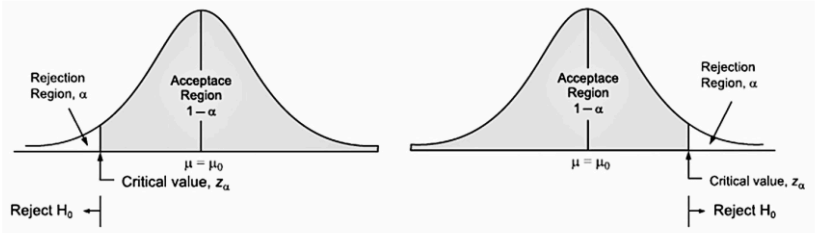


Fig. 5.2: (a) H_0: $\mu \geq \mu_0$; H1: $\mu \leq \mu_0$, Left-tailed Test **Fig. 5.2: (b)** H_0: $\mu \leq \mu_0$; H1: $\mu > \mu_0$, Right-tailed Test,

5.4 ERRORS IN HYPOTHESIS TESTING:

Ideally, the **hypothesis testing** procedure should lead to accepting the null hypothesis H0 when it is true and rejecting H_0 when it is not. However, the correct decision is not always possible.

An incorrect decision or error is possible since rejecting or accepting a hypothesis is based on sample data. A decision-maker may commit two types of errors while testing a null hypothesis.

The two kinds of mistakes that can be made in any hypothesis testing are shown in Table 5.2.
Hypothesis testing: Testing a statement or belief about a population parameter using information collected from a sample(s).

Table 5.2: Errors in Hypothesis Testing

Decision	*State of Nature*	
	H_0 is True	H_0 is False
Accept H_0	Correct decision with confidence(1-α)	Type II error(β)
Reject H_0	Type I error (α)	Correct decision(1-β)

Type I Error: This is the *probability of rejecting the null hypothesis when it is true*, and some alternative hypothesis is wrong. The likelihood of making a Type I error is denoted by the symbol α. The area represents it under the sampling distribution curve over the rejection region.

Type I error: The probability of rejecting a true null hypothesis.
The probability of making a Type I error is called the **level of significance.** The decision-maker decides the probability level of this error before the hypothesis test is performed. It is based on their tolerance for rejecting the true null hypothesis. The risk of making a Type I error depends on the cost and/or goodwill loss. The Type I error probability's complement (1–α) measures the probability level of not rejecting a true null hypothesis. It is also referred to as the *confidence level*.

Level of significance: The probability of rejecting a true null hypothesis due to sampling error.

Type II Error: This is the *probability of accepting the null hypothesis when it is false*, and some alternative hypothesis is true. The symbol denotes the likelihood of making a Type II β.

Type II error: The probability of accepting a false null hypothesis.
The probability of a Type II error varies with the actual values of the population parameter being tested when the null hypothesis H_0 is false. The likelihood of committing a Type II error depends on five factors: (i) the actual value of the population parameter being tested, (ii) the level of significance selected, (iii) the type of test (one or two-tailed tests) used to evaluate the null hypothesis, (iv) the sample standard deviation (also called standard error) and (v) the size of the sample.

A summary of specific critical values at various significance levels for the test statistic *z* is given in Table 5.3

Table 5.3: Summary of Certain Critical Values for Sample Statistic z

Rejection Region	Level of Significance			
	0.10	0.05	0.01	0.005
One-tailed region	±1.28	±1.645	±2.33	±2.58
Two-tailed region	±1.645	±1.96	±2.58	±2.81

5.4.1 Power of a Statistical Test
Another way of evaluating the goodness of a statistical test is to look at the complement of Type II error, which is stated as:
$1 - \beta = P$ (reject H_0 when H_1 is true)

The complement 1 – β of β, i.e., the probability of Type-II error, is the *Power of a statistical test* because it measures the probability of rejecting H_0 when it is true.

5.4.1.1 Power of a test: The ability (probability) of a test to reject the null hypothesis when it is false.
For example, suppose null and alternative hypotheses are stated as.
H_0: μ = 80 and H_1: μ = 80

When the null hypothesis is often false, another alternative value of the population mean, μ, is unknown. So, for each of the possible values of the population mean μ, the probability of committing a Type II error for several possible values of μ must be calculated.

Suppose a sample of size n = 50 is drawn from the given population to compute the probability of committing a Type II error for a specific alternative value of the population mean, μ. Let the sample mean so obtained be $\bar{x}$ = 71 with a standard deviation, s = 21. For the significance level, α = 0.05, and a two-tailed test, the table value of $z_{0.05}$ = ±1.96. But the deserved value from the sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

Since z_{cal} = – 3.03 value falls in the rejection region, the null hypothesis H_0 is rejected. The rejection of the null hypothesis leads to either making a correct decision or committing a Type II error. If the population mean is 75 instead of 80, then the probability of committing a Type II error is determined by computing a critical region for the mean $\bar{x}_c$. This value is used as the cutoff point between acceptance and rejection. If any sample mean obtained is less than (or greater than for right-tail rejection region), c, the null hypothesis is rejected. Solving for the critical value of the mean gives

$$z_c = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \text{ or } \pm 1.96 = \frac{\bar{x}_c - 80}{21/\sqrt{50}}$$

$$\bar{x}_c = 80 \pm 5.82 \text{ or } 74.18 \text{ to } 85.82$$

If μ = 75, then probability of accepting the false null hypothesis H_0 : μ = 80 when critical value is falling in the range $\bar{x}_c$ = 74.18 to 85.82 is calculated as follows:

$$z_1 = \frac{74.18 - 75}{21/\sqrt{50}} = -0.276$$

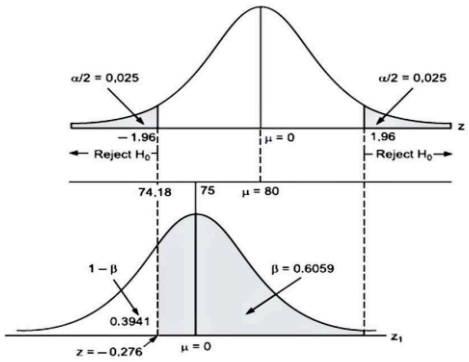
The area under the normal curve for z_1 = − 0.276 is 0.1064.

$$z_2 = \frac{85.82 - 75}{21\sqrt{50}} = 3.643$$

The area under the normal curve for z_2 = 3.643 is 0.4995
Thus, the probability of committing a Type II error (β) falls in the region:
β = P (74.18 < $\bar{x}_c$ < 85.82) = 0.1064 + 0.4995 = 0.6059

The total probability of 0.6059 of committing a Type II error (β) is the area to the right of $\bar{x}_c$ = 74.18 in the distribution. Hence, the power of the test is 1 − β = 1 − 0.6059 = 0.3941, as shown in Fig. 5.3

Figure 5.3 (a) Sampling distribution with $H_0 : \mu = 80$ **Figure 5.3 (b)** Sampling distribution with $H_0 : \mu = 75$



The decision to keep α or β low depends on which type of error is more costly. However, if both mistakes are expensive, keeping both α and β low can make inferences more reliable by reducing the variability of observations. It is preferred to have a large sample size and a low α value.

Few relations between two errors α and β, the Power of test 1 − β, and the sample size *n* are stated below:

1. If α (the sum of the two tail areas in the curve) is increased, the shaded area corresponding to β gets smaller, and vice versa.
2. The β value can be increased for a fixed α by increasing the sample size *n*.

Special Case: Suppose hypotheses are defined as:
H_0: u. = 80 and H_1: u. ≤ 80
Given n = 50, s = 21 and $\bar{x}$ = 71. For α = 0.05 and a left-tailed test, the table value $z_{0.05}$ = −1.645. The observed z value from the sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

The critical value of the sample mean $\bar{x}_c$ for a given population mean μ = 80 is given by:

$$z_c = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \text{ or } -1.645 = \frac{\bar{x}_c - 80}{21/\sqrt{50}}$$

$$\bar{x}_c = 75.115$$

Figure 5.4 (a) Sampling distribution with H_0: $\mu = 80$ **Figure 5.4 (b)** Sampling distribution with H_0: $\mu = 78$

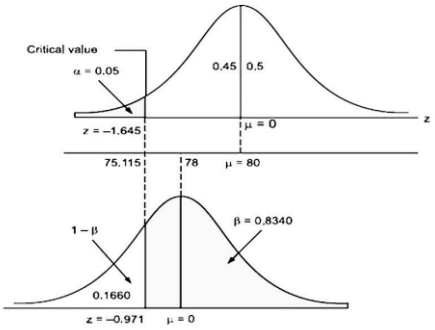


Figure 5.4(a) shows that the distribution of values contains the critical value of mean c = 75.115 and below, which means H_0 will be rejected. Figure 5.4(b) shows the distribution of values when the alternative population mean value $\mu = 78$ is actual. If H_0 is false, it is impossible to reject the null hypothesis H0 whenever the sample mean is in the acceptance region, $\bar{x} \geq 75.151$. Thus, the critical value is computed by extending it and solving for the area to the right of $\bar{x}_c$ as follows:

$$z_1 = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} = \frac{75.115 - 78}{21/\sqrt{50}} = -0.971$$

This value of z yields an area of 0.3340 under the standard curve. Thus, the probability = 0.3340 + 0.5000 = 0.8340 of committing a Type II error is all the area to the right of $\bar{x}_c = 75.115$.

Remark: In general, if the alternative value of the population mean μ is relatively greater than its hypothesized value, then the probability of committing a Type II error is negligible compared to when the alternative value is close to the hypothesized value. The likelihood of committing a Type II error decreases as the alternative values are more significant than the hypothesized mean of the population.

5.5 SUMMARY:

- Hypothesis testing involves making assumptions (hypotheses) about a population parameter and using sample data to evaluate them.
- Hypotheses are commonly stated in null (H_0) and alternative (H_1) formats.
- The rationale behind hypothesis testing is to assess whether observed sample results are due to chance.
- The general Procedure includes setting hypotheses, choosing a significance level (α), computing a test statistic, and deciding.
- Tests can be one-tailed or two-tailed, depending on the direction of the hypothesis.

5.6 TECHNICAL TERMS:

- **Hypothesis:** A statement or assumption about a population parameter that can be tested using sample data.
- **Null Hypothesis (H_0):** A statement that there is no effect or difference; it is tested for possible rejection.
- **Alternative Hypothesis (H_1 or Ha):** A statement that contradicts the null hypothesis, indicating the presence of an effect or difference.
- **Significance Level (α):** The probability of rejecting the null hypothesis when it is true, commonly set at 0.05 or 5%.
- **Test Statistic**: A standardized value calculated from sample data used to decide whether to reject H_0.
- **Critical Region:** The range of values of the test statistic that leads to rejection of the null hypothesis.
- **Type I Error:** Occurs when the null hypothesis is wrongly rejected (false positive).
- **Type II Error** happens when the null hypothesis is wrongly accepted (false negative).
- **Power of a Test:** The probability of correctly rejecting a false null hypothesis; calculated as $1 - \beta$.
- **One-tailed Test:** A test that checks for a deviation in one specific direction (greater than or less than).
- **Two-tailed Test:** A test that checks for a deviation in both directions (either greater or less).
- **P-value:** The probability of obtaining the observed results, or more extreme, if the null hypothesis is true.
- **Sampling Distribution:** The probability distribution of a statistic based on a random sample.
- **Statistical Decision:** The conclusion drawn from a hypothesis test—either to reject or fail to reject H_0.

5.7 SELF-ASSESSMENT QUESTIONS:

1. Describe the various steps involved in testing a hypothesis. What is the role of standard error in testing a hypothesis?
2. What do you understand about the null hypothesis and its significance level? Explain with the help of one example.
3. What is a test statistic? How is it used in hypothesis testing?
4. Define the term ‘level of significance’. How is it related to the probability of committing a Type I error?

a. Explain the general steps needed to carry out a test of any hypothesis.
b. Explain clearly the Procedure for testing the hypothesis. Also, point out the assumptions in hypothesis testing in large samples.

5. This is always a trade-off between Type I and Type II errors. Discuss.
6. Define the standard error of a statistic. How is it helpful in testing hypotheses and decision-making?
7. Define the terms 'decision rule' and 'critical value.' What is the relationship between these terms?
8. Write short notes on the following:
    1. Acceptance and rejection regions
    2. Type I and Type II errors
    3. Null and alternative hypotheses
    4. One-tailed and two-tailed tests

5.8 SUGGESTED READINGS:

1. Gupta, S. C., & Gupta, I. (2023). *Business Statistics* (18th Revised ed.). Himalaya Publishing House.
2. Sharma, J.K. (2020). *Business Statistics*. Pearson Education India.
3. Bajpai, N. (2019). *Business Statistics* (2nd ed.). Pearson Education.
4. Hooda, R. P. (2014). *Statistics for Business and Economics* (4th ed.). Macmillan Publishers India.

Dr. G. Malathi

LESSON- 6

HYPOTHESIS TESTING FOR POPULATION PARAMETERS WITH LARGE SAMPLES

OBJECTIVES:

The purpose of studying this lesson is:

- To understand the concept of hypothesis testing for a single population mean.
- To explore the connection between confidence intervals and hypothesis testing.
- To learn how to use the p-value approach in testing hypotheses.
- To compare means from two different populations using statistical tests.
- To apply appropriate test statistics based on sample sizes and population variance knowledge.
- To interpret results to make data-driven decisions.

STRUCTURE:

6.1 HYPOTHESIS TESTING FOR SINGLE POPULATION MEAN:

Hypothesis testing involving large samples ($n \geq 30$) is based on the assumption that the population from which the sample is drawn has a normal distribution. Consequently, the sampling distribution of the mean $\bar{x}$ is also standard. Even if the population does not have a normal distribution, the sampling distribution of the mean $\bar{x}$ is assumed to be normal due to the central limit theorem because the sample size is large.

Two-tailed Test: Let μ_0 be the hypothesized value of the population mean to be tested. For this, the null and alternative hypotheses for a two-tailed test are defined as:

$$H_0 : \mu = \mu_0 \quad \text{or} \quad \mu - \mu_0 = 0$$

and

$$H_1 : \mu \neq \mu_0$$

If the standard deviation σ of the population is known, then, based on the central limit theorem, the sampling distribution of the mean $\bar{x}$ will follow the standard normal distribution for a large sample size. The z-test statistic is given by

$$\text{Test-statistic: } z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

In this formula, the numerator $\bar{x} - \mu$, measures how far (in an absolute sense) the observed sample mean $\bar{x}$ is from the hypothesized mean μ. The denominator $\sigma_{\bar{x}}$ is the *standard error of the mean*, so the *z*-test statistic represents how many standard errors $\bar{x}$ are from μ.
If the population standard deviation σ is unknown, then a sample standard deviation *s* is used to estimate σ. The value of the *z*-test statistic is given by

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The two rejection areas in a two-tailed test are determined so that half the level of significance, a/2, appears in each tail of the distribution of the mean. Hence, za/2 represents the standardized normal variate corresponding to a/2 in both tails of the normal curve. The decision rule based on the sample mean for the two-tailed test takes the form:

- Reject H0 if $z_{cal} \leq -z_{\alpha/2}$ or $z_{cal} \geq z_{\alpha/2}$
- Accept H0 if $-z_{\alpha/2} < z < z_{\alpha/2}$

where $z_{\alpha/2}$ is the table value (also called CV, critical value) of *z* at a chosen significance α.

Left-tailed Test Large sample ($n > 30$) hypothesis testing about a population mean for a left-tailed test is of the form

$$H_0 : \mu \geq \mu_0 \quad \text{and} \quad H_1 : \mu < \mu_0$$

$$\text{Test statistic: } z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Decision rule:
- Reject H_0 if $z_{cal} \leq -z_{\alpha}$ (Table value of z at α)
- Accept H_0 if $z_{\alpha/2}\ z > -z_{\alpha}$

Right-tailed Test Large sample ($n > 30$) hypothesis testing about a population mean for a right-tailed test is of the form

$$H_0 : \mu \leq \mu_0 \text{ and } H_1 : \mu > \mu_0 \text{ (Right-tailed test)}$$

$$\text{Test statistic: } z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Decision rule:
- Reject H_0 if $z_{cal} \geq z_{\alpha}$ (Table value of z at α)
- Accept H_0 if $z_{cal} < z_{\alpha}$

6.1.1 Relationship between Interval Estimation and Hypothesis Testing

Consider the following statements of null and alternative hypotheses:

- $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ (Two-tailed test)
- $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ (Right-tailed test)
- $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$ (Left-tailed test)

The following are the confidence intervals in all three above cases where the hypothesized value μ0 of the population mean μ is likely to fall. Accordingly, the decision to accept or reject the null hypothesis will be taken.

Two-tailed test: Two critical values, CV_1 and CV_2, one for each tail of the sampling distribution, are computed as follows:

Two-tailed test: The test of a null hypothesis can be rejected when the sample statistic is at either end of the sampling distribution.

(a) Known σ	(b) Unknown σ
Normal population : Any sample size, n Any population : Large sample size n	Any population : Large sample size, n
$CV_1 = \mu_0 - z_{\alpha/2}\,\sigma_{\bar{x}}$ $CV_2 = \mu_0 + z_{\alpha/2}\sigma_{\bar{x}}$	$CV_1 = \mu_0 - z_{\alpha/2}\,s_{\bar{x}}$ $CV_2 = \mu_0 + z_{\alpha/2}\,s_{\bar{x}}$
where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$	$s_{\bar{x}} = s/\sqrt{n}$

Decision rule:
- Reject H_0 when $\bar{x} \le CV_1$ or $\bar{x} \ge CV_2$.
- Accept H_0 when $CV_1 < \bar{x} < CV_2$

Left-tailed test The critical value for the left tail of the sampling distribution is computed as follows:

(a) Known σ	(b) Unknown σ
Normal population : Any sample size, n Any population : Large sample size, n	Any population : Large sample size, n
$CV = \mu_0 - z_\alpha\,\sigma_{\bar{x}}$	$CV = \mu_0 - z_\alpha\,s_{\bar{x}}$

Decision rule:
- Reject H_0 when $\bar{x} \le CV$
- Accept H_0 when $\bar{x} > CV$

Right-tailed test: The critical value for the right tail of the sampling distribution is computed as follows:

(a) Known σ	(b) Unknown σ
Normal population : Any sample size, n Any population : Large sample size, n	Any population : Large sample size, n
$CV = \mu_0 + z_\alpha\,\sigma_{\bar{x}}$	$CV = \mu_0 + z_\alpha\,s_{\bar{x}}$

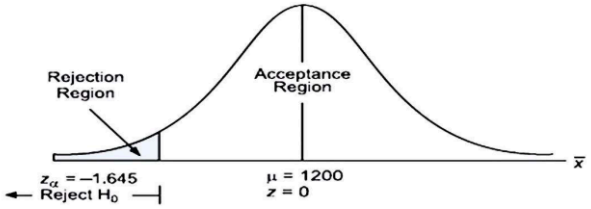
Decision rule:
- Reject H_0 when $\bar{x} \ge CV$
- Accept H_0 when $\bar{x} < CV$

Example 6.1: Individual filing of income tax returns before 30 June had an average refund of Rs 1200. Consider the population of 'last-minute' filers who file their returns during the last week of June. For a random sample of 400 individuals who filed a return between 25 and 30 June, the sample mean refund was Rs 1054, and the sample standard deviation was Rs 1600. Using a 5 percent level of significance, test the belief that the individuals who wait until the last week of June to file their returns get a higher refund than early filers.

Solution: Since the population standard deviation is not given, the standard error must be estimated with $s_{\bar{x}}$. Let us take the null hypothesis H_0 that the individuals who wait until the last week of June to file their returns get a higher return than the early filers that is,

H_0: $\mu \geq 1200$ and H_1: $\mu < 1200$ (Left-tailed test)

Figure 6.1



Given, $n = 400$, $s = 1600$, $\bar{x} = 1054$, $\alpha = 5\%$. Thus using the z-test statistic

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1054 - 1200}{1600/\sqrt{400}} = -\frac{146}{80} = -1.825$$

Since the calculated value $z_{cal} = -1.825$ is less than its critical value $z_\alpha = -1.645$ at $\alpha = 0.05$ significance level, the null hypothesis, H0, is rejected, as shown in Fig 6.1. Hence, we conclude that individuals who wait until the last week of June will likely receive a refund of less than Rs 1200.

Alternative approach: $CV = \mu_0 - z_\alpha \sigma_{\bar{x}} = 1200 - 1.645 \times (1600/\sqrt{400})$

$= 1200 - 131.6 = 1068.4$

Since $\bar{x}\,(= 1054) < CV(= 1068.4)$, the null hypothesis H_0 is rejected

Example 6.2: A packaging device is set to fill detergent powder packets with a mean weight of 5 kg, with a standard deviation of 0.21 kg. The weight of packets can be assumed to be normally distributed. The weight of packets is known to drift upwards over time due to machine fault, which is not tolerable. A random sample of 100 packets is taken and weighed. This sample has a mean weight of 5.03 kg. Can we conclude that the mean weight produced by the machine has increased? Use a 5 percent level of significance.

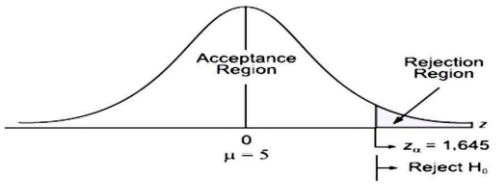
Solution: Let us take the null hypothesis H_0 that the mean weight has increased, that is,

$$H_0 : \mu \geq 5 \text{ and } H_1 : \mu < 5$$

Given $n = 100$, $\bar{x} = 5.03$ kg, $\sigma = 0.21$ kg and $\alpha = 5$ per cent. Thus using the z-test statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{5.03 - 5}{0.21\sqrt{100}} = \frac{0.03}{0.021} = 1.428$$

Figure 6.2



Since the calculated value zcal = 1.428 is less than its critical value zα = 1.645 at α = 0.05, the null hypothesis, H0, is accepted, as shown in Fig. 6.2. Hence, we conclude that the mean weight is likely to be more than 5 kg.

Alternative approach: $CV = \mu_0 + z_\alpha \sigma_{\bar{x}} = 5 + 1.645 \times (0.21/\sqrt{100})$

$= 5 + 0.034 = 5.034$

Since $\bar{x}\,(= 5.03) < CV(= 5.034)$, H_0 is accepted.

Example 6.3: The mean lifetime of a sample of 400 fluorescent light bulbs produced by a company is 1600 hours with a standard deviation of 150 hours. Test the hypothesis that the mean lifetime of the bulbs produced in general is higher than the mean life of 1570 hours at α = 0.01 level of significance.

Solution: Let us take the null hypothesis that the mean lifetime of bulbs is not more than 1570 hours, that is

$$H_0 : \mu \leq 1570 \text{ and } H_1 : \mu > 1570 \quad \text{(Right-tailed test)}$$

Given $n = 400$, $\bar{x} = 1600$ hours, $s = 150$ hrs and $\alpha = 0.01$. Thus using the z-test statistic.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1600 - 1570}{150/\sqrt{400}} = \frac{30}{7.5} = 4$$

Since the calculated value $z_{cal} = 4$ is greater than its critical value $z\alpha = \pm 2.33$, the H0 is rejected. Hence, we conclude that the company's bulbs may have a mean lifetime of more than 1570 hours.

Alternatively approach: $CV = \mu_0 + z_\alpha s_{\bar{x}} = 1570 + 2.33 \times (150/\sqrt{400})$

$= 1570 + 17.475 = 1587.475$

Since $\bar{x}(= 1600) > CV (= 1587.47)$, the null hypothesis H_0 is rejected.

Example 6.4: A continuous manufacturing process of steel rods is said to be in a 'state of control' and produces acceptable rods if the mean diameter of all rods produced is 2 inches. Although the process standard deviation exhibits stability over time, with a standard deviation of O = 0.01 inch. The process mean may vary due to operator error or problems with process adjustment. Periodically, random samples of 100 rods are selected to determine whether the process is producing acceptable rods. If the result of a test indicates that the process is out of control, it is stopped, and the source of trouble is sought. Otherwise, it is allowed to continue operating. A random sample of 100 rods is selected, resulting in a mean of 2.1 inches. Test the hypothesis to determine whether the process should be continued.

Solution: Since rods that are either too narrow or too wide are unacceptable, the low and high values of the sample mean lead to the rejection of the null hypothesis. Consider the null hypothesis H_0, which states that the process may continue when the diameter is 2 inches. Consequently, the rejection region is on both tails of the sampling distribution. The null and alternative hypotheses are stated as follows:

H_0: $\mu = 2$ inches (continue the process)
H_1: $\mu \neq 2$ inches (stop the process)

Given $n = 100$, $\bar{x} = 2.1$, $\sigma = 0.01$, $\alpha = 0.01$. Using the z-test statistic

$$z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{2.1-2}{0.01/\sqrt{100}} = \frac{0.1}{0.001} = 100$$

Since $z_{cal} = 100$ value is more than its critical value $z_{\alpha/2} = 2.58$ at a = 0.01, the null hypothesis, H_0, is rejected. Thus, the process of determining the source of trouble must be stopped.

Alternative approach:

$$CV_1 = \mu_0 - z_{\alpha/2}\,\sigma_{\bar{x}} = \mu_0 - z_{\alpha/2}(\sigma/\sqrt{n})$$
$$= 2 - 2.58\times(0.01/\sqrt{100}) = 2 - 0.003 = 1.997$$
$$CV_2 = \mu_0 + z_{\alpha/2}\,\sigma_{\bar{x}} = 2 + 2.58\times(0.01/\sqrt{100})$$
$$= 2+0.003 = 2.003$$

Since $\bar{x}\,(= 2.1) \geq CV_2 (= 2.003)$, the null hypothesis is rejected.

Example 6.5: An ambulance service claims that it takes, on average, 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency that licenses ambulance services then timed 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.8 minutes. Does this constitute evidence that the figure claimed is too low at the 1 percent significance level?

Solution: Let us consider the null hypothesis H_0, which states that 'the claim is the same as observed' and that the alternative hypothesis is 'the claim is different from what is observed.' These two hypotheses are written as:

$$H_0 : \mu = 8.9 \text{ and } H_1 : \mu \neq 8.9$$

Given $n = 50$, $\bar{x} = 9.3$, and $s = 1.8$. Using the z-test statistic, we get

$$z = \frac{\bar{x}-\mu}{s_{\bar{x}}} = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{9.3-8.9}{1.8/\sqrt{50}} = \frac{0.4}{0.254} = 1.574$$

The null hypothesis is accepted since $z_{cal} = 1.574$ is less than its critical value $z_{\alpha/2} = \pm 2.58$, at $\alpha = 0.01$. Thus, there is no difference between the average time observed and claimed.

6.1.2 *p*- Value Approach to Test Hypothesis of Single Population Mean

The ***p*-value** is another approach for hypothesis testing population mean based on a large sample. This is often referred to as the *observed significance level*, the smallest significance level for which null hypothesis H_0 can be rejected. It is the risk of committing a Type I error when the null hypothesis, H_0, is rejected based on the observed value of the test statistic. The *p*-value measures the strength of evidence against H_0, i.e., a *p*-value is a way to express the likelihood that H_0 is not true. In other words, the *p-value is the probability of observing a sample value as extreme as or more extreme than the value of the test statistic, given that the null hypothesis H_0 is true.* The advantage of this approach is that the *p*-value can be compared directly to the level of significance α.

P-value: The probability of getting the sample statistic or a more extreme value when the null hypothesis is true.

The decision rules for accepting or rejecting a null hypothesis based on the *p*-value are as follows:

1. For a left-tailed test, the *p*-value is the area to the left of the calculated value of the test statistic. For instance, if $z_{cal} = -1.76$, then the area to the left of it is $0.5000 - 0.4608 = 0.0392$, or the *p* -p-value is 3.92 percent.

2. For the right-tailed test, the *p*-value is the area to the right of the calculated value of the test statistic. For instance, if $_{scal}e = +2.00$, then the area to the right of it is 0.5000 - 0.4772 = 0.0228, or the *p*-value is 2.28 percent.

Thus, the decision rules for left and right-tailed tests are as follows.

- Reject H_0 if *p*-value $\leq \alpha$
- Accept H_0 if *p*-value $> \alpha$

3. The p-value is twice the tail area for a two-tailed test. If the calculated value of the test statistic falls in the left tail (or right tail), then the area to the left (or right) of the calculated value is multiplied by 2.

Example 6.6: An auto company decided to introduce a new six-cylinder car whose mean petrol consumption is claimed to be lower than that of the existing auto engine. The mean petrol consumption for 50 cars was 10 km per liter, with a standard deviation of 3.5 km per liter. Test for the company at the 5 percent level of significance, the claim that in the new car, petrol consumption is 9.5 km per liter on average.

Solution: Let us assume the null hypothesis H_0 that there is no significant difference between the company's claim and the sample average value, that is,

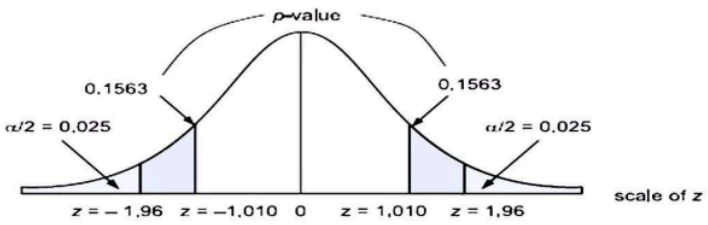
H_0: $\mu = 9.5$ km/litre and H_1: $\mu \neq 9.5$ km/litre

Given $\bar{x} = 10$, $n = 50$, $s = 3.5$, and $z_{\alpha/2} = 1.96$ at $\alpha = 0.05$ level of significance. Thus, using the *z*-test statistic

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{10 - 9.5}{3.5/\sqrt{50}} = 1.010$$

The null hypothesis is accepted since zcal = 1.010 is less than its critical value zα/2 = 1.96 at α = 0.05 significance level. Hence, we can conclude that the new car's petrol consumption is 9.5 km/liter.

Figure 6.3



The *p*-value approach: The null hypothesis, H0, is accepted because z_{cal} = 1.010 lies in the acceptance region. The probability of finding z_{cal} = 1.010 or more is 0.3437 (from the regular table). The *p-value is the area to the right and left of the calculated value of the z-test statistic (for the* two-tailed test). Since z_{cal} = 1.010, the area to its right is 0.5000 − 0.3437 = 0.1563, as shown in Fig. 6.3.
Since it is a two-tailed test, the *p*-value becomes 2(0.1563) =0.3126. Since 0.3126 > α=0.05, the null hypothesis H_0 is accepted.

6.2 HYPOTHESIS TESTING FOR DIFFERENCE BETWEEN TWO POPULATION MEANS:

If we have two independent populations, each having its mean and standard deviation as:

Population	*Mean*	*Standard Deviation*
1	μ_1	σ_1
2	μ_2	σ_2

Then, we can extend the hypothesis testing concepts developed in the previous section to test whether there is any significant difference between the means of these populations.
Let two independent random samples of large size, n1, and n_2, be drawn from the first and second populations, respectively. Let the sample means be $\bar{x}_1$ and $\bar{x}_2$.

The *z*-test statistic used to determine the difference between the population means ($\mu_1 - \mu_2$) is based on the difference between the sample means ($\bar{x}_1 - \bar{x}_2$) because the sampling distribution of $\bar{x}_1 - \bar{x}_2$ has the property E ($\bar{x}_1 - \bar{x}_2$) = ($\mu_1 - \mu_2$). This test statistic will follow the standard normal distribution for a large sample due to the central limit theorem. The *z*-test statistic is

$$\text{Test statistic: } z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\sigma_{\bar{x}_1 - \bar{x}_2}$ = standard error of the statistic $(\bar{x}_1 - \bar{x}_2)$
$\bar{x}_1 - \bar{x}_2$ = difference between two sample means, that is, sample statistic
$\mu_1 - \mu_2$ = difference between population means, that is, hypothesized population parameter

If $\sigma_1^2 = \sigma_2^2$, the above formula algebrically reduces to:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Suppose the standard deviations σ1 and σ2 of each population are unknown. In that case, we may estimate the standard error of the sampling distribution of the sample statistic $\bar{x}_1 - \bar{x}_2$ by substituting the sample standard deviations s_1 and s_2 as estimates of the population standard deviations. Under this condition, the standard error of $\bar{x}_1 - \bar{x}_2$ is estimated as:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The standard error of the *difference between the standard deviation of a sampling distribution* is given by

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

The null and alternative hypotheses are stated as follows:

Null hypothesis : $H_0 : \mu_1 - \mu_2 = d_0$
Alternative hypothesis :

One-tailed Test	*Two-tailed Test*
$H_1 : (\mu_1 - \mu_2) > d_0$	$H_1 : (\mu_1 - \mu_2) \neq d_0$
$H_1 : (\mu_1 - \mu_2) < d_0$	

Where d_0 is some specified difference that is desired to be tested. If there is no difference between μ_1 and μ_2, i.e. $\mu_1 = \mu_2$, then $d_0 = 0$.

Decision rule: Reject H_0 at a specified level when:
One-tailed test: $Z_{cal} > Z_\alpha$ or p-value < α
Two-tailed test: $Z_{cal} > Z_{\alpha/2}$

Example 6.7: A firm believes that the tires produced by process A last longer than those produced by process B on average. To test this belief, random samples of tires made by the two methods were tested, and the results are as follows:

Process	Sample Size	Average Lifetime (in km)	Standard Deviation (in km)
A	50	22,400	1000
B	50	21,800	1000

Is there evidence at a 5 percent significance level that the firm is correct in its belief?
Solution: Let us take the null hypothesis that there is no significant difference in the average life of tires produced by processes A and B, that is,

$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2$$

Given, $\bar{x}_1 = 22{,}400$ km, $\bar{x}_2 = 21{,}800$ km, $\sigma_1 = \sigma_2 = 1000$ km, and $n_1 = n_2 = 50$. Thus using the z-test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{22{,}400 - 21{,}800}{\sqrt{\frac{(1000)^2}{50} + \frac{(1000)^2}{50}}} = \frac{600}{\sqrt{20{,}000 + 20{,}000}} = \frac{600}{200} = 3$$

Since the calculated value $z_{cal} = 3$ is more than its critical value $z_{\alpha/2} = \pm 1.645$ at $\alpha = 0.05$ significance level, H_0 is rejected. Hence, we can conclude that the tires produced by process A last longer than those produced by process B.

The p-value approach:

$$p\text{-value} = P(z > 3.00) + P(z < -3.00) = 2\,P(z > 3.00)$$
$$= 2(0.5000 - 0.4987) = 0.0026$$

Since *the p-value of 0.026 is less than the* specified significance level $\alpha = 0.05$, H_0 is rejected.

Example 6.8: An experiment was conducted to compare the mean time in days required to recover from a common cold for persons given a daily dose of 4 mg of vitamin C versus those not given a vitamin supplement. Suppose that 35 adults were randomly selected for each treatment category and that the mean recovery times and standard deviations for the two groups were as follows:

	Vitamin C	No Vitamin Supplement
Sample Size	35	35
Sample Mean	5.8	6.9
Sample Standard Deviation	1.2	2.9

Test the hypothesis that vitamin C reduces the mean time required to recover from a common cold and its complications at the significance level $\alpha = 0.05$.
Solution: Let us take the null hypothesis that the use of vitamin C reduces the mean time required to recover from the common cold, that is

$$H_0 : (\mu_1 - \mu_2) \le 0 \text{ and } H_1 : (\mu_1 - \mu_2) > 0$$

Given $n_1 = 35$, $\bar{x}_1 = 5.8$, $s_1 = 1.2$ and $n_2 = 35$, $\bar{x}_2 = 6.9$, $s_2 = 2.9$. The level of significance, $\alpha = 0.05$. Substituting these values into the formula for z-test statistic, we get

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{5.8 - 6.9}{\sqrt{\frac{(1.2)^2}{35} + \frac{(2.9)^2}{35}}} = \frac{-1.1}{\sqrt{0.041 + 0.240}} = -\frac{1.1}{0.530} = -2.605$$

Using a one-tailed test with significance level $\alpha = 0.05$, the critical value is $z_\alpha = 1.645$. Since $z_{cal} < z_\alpha$ (= 1.645), the null hypothesis H_0 is rejected. Hence, we can conclude that vitamin C does not reduce the time required to recover from the common cold.

Example 6.9: The Educational Testing Service conducted a study to investigate the difference between the scores of female and male students on the Mathematics Aptitude Test. The study identified a random sample of 562 female and 852 male students who had achieved the same high score on the mathematics portion of the test. The female and male students were viewed as having similar high mathematics abilities. The verbal scores for the two samples are given below:

	Female	Male
Sample Mean	547	525
Sample Standard Deviation	83	78

Do the data support the conclusion that given populations of female and male students with similar high abilities in mathematics, the female students will have a significantly higher verbal ability? Test at $\alpha = 0.05$ significance level. What is your conclusion?

Solution: Let us take the null hypothesis that the female students have a high level of verbal ability, that is,

$$H_0 : (\mu_1 - \mu_2) \geq 0 \text{ and } H_1 : (\mu_1 - \mu_2) < 0$$

Given, for female students: $n_1 = 562$, $\bar{x}_1 = 547$, $s_1 = 83$, for male students: $n_2 = 852$, $\bar{x}_2 = 525$, $s_2 = 78$, and $\alpha = 0.05$.

Substituting these values into the z-test statistic, we get

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{547 - 525}{\sqrt{\frac{(83)^2}{562} + \frac{(78)^2}{852}}}$$

$$= \frac{22}{\sqrt{12.258 + 7.140}} = \frac{22}{\sqrt{19.398}} = \frac{22}{4.404} = 4.995$$

Using a one-tailed test with $\alpha = 0.05$ significance level, the critical value of a z-test statistic is $z_\alpha = \pm 1.645$. Since $z_{cal} = 4.995$ is more than the critical value $z_\alpha = 1.645$, the null hypothesis, H0, is rejected. Hence, we conclude that there is no insufficient evidence to declare that the difference between the verbal ability of female and male students is significant.

Example 6.10: In a sample of 1000, the mean is 17.5, and the standard deviation is 2.5. In another sample of 800, the mean is 18, and the standard deviation is 2.7. Assuming that the samples are independent, discuss whether the two samples could have come from a population with the same standard deviation.

Solution: Let us take the hypothesis that there is no significant difference in the standard deviations of the two samples, that is, H_0: $\sigma_1 = \sigma_2$ and H_1: $\sigma_1 \neq \sigma_2$.

Given, $\sigma_1 = 2.5$, $n_1 = 1000$ and $\sigma_2 = 2.7$, $n_2 = 800$. Thus, we have

$$\text{Standard error, } \sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_1}}$$

$$= \sqrt{\frac{(2.5)^2}{2000} + \frac{(2.7)^2}{1600}} = \sqrt{\frac{6.25}{2000} + \frac{7.29}{1600}} = 0.0876$$

Applying the z-test statistic, we have

$$z = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{2.7 - 2.5}{0.0876} = \frac{0.2}{0.0876} = 2.283$$

Since zcal = 2.283 is more than its critical value z = 1.96 at α = 5 percent, the null hypothesis H0 is rejected. Hence, we conclude that the two samples did not come from a population with the same standard deviation.

Example 6.11: The mean wheat production from a sample of 100 fields is 200 lbs per acre with a standard deviation of 10 lbs. Another sample of 150 fields gives the mean at 220 lbs per

acre with a standard deviation of 12 lbs. Assuming the standard deviation of the universe as 11 lbs, find at a 1 percent level of significance whether the two results are consistent.

Solution: Let us take the hypothesis that the two results are consistent, that is

$$H_0 : \sigma_1 = \sigma_2 \text{ and } H_1 : \sigma_1 \neq \sigma_2.$$

Given $\sigma_1 = \sigma_2 = 11$, $n_1 = 100$, $n_2 = 150$. Thus

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma^2}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{(11)^2}{2}\left(\frac{1}{100} + \frac{1}{150}\right)} = 1.004$$

Applying the z-test statistic we have

$$z = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{10 - 12}{1.004} = -\frac{2}{1.004} = -1.992$$

The null hypothesis is accepted since the zcal = – 1.992 is more than its critical value z = – 2.58 at $\alpha = 0.01$. Hence, we conclude that the two results are likely to be consistent.

6.3 SUMMARY:

Hypothesis testing is a statistical method for inferring population parameters. It involves setting up null and alternative hypotheses and evaluating them using sample data. Interval estimation provides a range of plausible values for the parameter and is closely related to hypothesis testing. The p-value approach quantifies the strength of evidence against the null hypothesis. Testing the difference between two population means helps determine if the two groups differ significantly. These methods support informed decisions in scientific and business applications.

6.4 TECHNICAL TERMS:

- **Hypothesis Testing:** A procedure to decide whether to accept or reject a statistical hypothesis based on sample data.
- **Interval Estimation:** Provides a range (confidence interval) likely to contain the population parameter.
- **P-Value:** The probability of obtaining results at least as extreme as the observed ones, assuming the null hypothesis is true.
- **Null Hypothesis (H_0):** An assumption that no effect or difference exists.
- **Alternative Hypothesis (H_1):** The statement we seek evidence for, indicating a significant effect or difference.
- **Confidence Interval:** A range of values derived from sample data within which the actual population parameter is expected to lie with a certain confidence level.
- **Test Statistic:** A standardized value used to decide whether to reject the null hypothesis.
- **Significance Level (α):** The threshold probability for rejecting the null hypothesis (commonly 0.05).

6.5 SELF-ASSESSMENT QUESTIONS:

1. The mean breaking strength of the cables supplied by a manufacturer is 1800, with a standard deviation of 100. A new technique in the manufacturing process is claimed to have increased the breaking strength of the cables. A sample of 50 cables was tested to test this claim. The mean breaking strength is 1850. Can we support the claim at a 0.01 level of significance?

2. A sample of 100 households in a village was taken, and the average income was found to be Rs 628 per month with a standard deviation of Rs 60 per month. Find the standard error of the mean and determine 99 percent confidence limits within which the income of all the people in this village is expected to lie. Also, test the claim that the average income was Rs 640 per month.
3. A random sample of boots worn by 40 combat soldiers in a desert region showed an average life of 1.08 years with a standard deviation of 0.05. Under the standard conditions, the shoes are known to have an average life of 1.28 years. Is there a reason to assert at a level of significance of 0.05 that use in the desert causes the mean life of such boots to decrease?
4. An ambulance service claims it takes 8.9 minutes to reach its destination in emergency calls. To check this claim, the agency that licenses ambulance services had them timed on 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.8 minutes. At the level of significance of 0.05, does this constitute evidence that the figure claimed is too low?
5. A simple sample of the heights of 6400 Englishmen has a mean of 67.85 inches and a standard deviation of 2.56 inches. In contrast, a simple sample of heights of 1600 Austrians has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the Austrians are taller than the Englishmen on average? Give reasons for your answer.
6. A man buys 50 electric bulbs of 'Philips' and 50 electric bulbs of 'HMT'. He finds that 'Philips' bulbs gave an average life of 1500 hours with a standard deviation of 60 hours, and 'HMT' bulbs gave an average life of 1512 hours with a standard deviation of 80 hours. Is there a significant difference in the mean life of the two kinds of bulbs?

6.6 SUGGESTED READINGS:

1. Gupta, S. C., & Gupta, I. (2023). *Business Statistics* (18th Revised ed.). Himalaya Publishing House.
2. Bajpai, N. (2019). *Business Statistics* (2nd ed.). Pearson Education.
3. Sharma, J.K. (2020). *Business Statistics*. Pearson Education India.
4. Hooda, R. P. (2014). *Statistics for Business and Economics* (4th ed.). Macmillan Publishers India.

Dr. G. Malathi

LESSON- 7

HYPOTHESIS TESTING FOR SINGLE-SAMPLE PROPORTION & T-TEST

OBJECTIVES:

The purpose of studying this lesson is:

1. To understand hypothesis testing procedures for a single sample proportion.
2. To learn how to apply the t-test for small sample sizes when testing population means.
3. To explore the properties and significance of the Student's t-distribution.
4. To identify scenarios where the t-distribution is applicable in hypothesis testing.
5. To compare two populations means using both independent and paired t-tests.

STRUCTURE:

Sometimes, instead of testing a hypothesis about a population mean, a population proportion (a fraction, ratio, or percentage) p of values that indicates the part of the population or sample having a particular attribute of interest is considered. For this, a random sample of size n is selected to compute the proportion of elements having a specific attribute of interest (also called success) in it as follows:

$$\bar{p} = \frac{\text{Number of successes in the sample}}{\text{Sample size}} = \frac{x}{n}$$

The value of this statistic is compared with a hypothesized population proportion p_0 to test the hypothesis.

The three forms of null hypothesis and alternative hypothesis about the hypothesized population proportion p are as follows:

Null hypothesis	*Alternative hypothesis*
• $H_0 : p = p_0$	$H_1 : p \neq p_0$ (Two-tailed test)
• $H_0 : p \geq p_0$	$H_1 : p < p_0$ (Left-tailed test)
• $H_0 : p \leq p_0$	$H_1 : p > p_0$ (Right-tailed test)

7.1 HYPOTHESIS TESTING FOR SINGLE SAMPLE PROPORTION:

To test a hypothesis, it is assumed that the sampling distribution of a proportion follows a standardized normal distribution. Then, using the value of the sample proportion $\bar{p}$ and its standard deviation $\sigma_{\bar{p}}$, we compute a value for the z-test statistic as follows:

$$\text{Test statistic } z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

The comparison of the z-test statistic value to its critical (table) value at a given level of significance enables us to test the null hypothesis about a population proportion based on the difference between the sample proportion and the hypothesized population proportion.

Decision rule: Reject H_0 at a specified level when:
One-tailed test: $Z_{cal} > Z_\alpha$ or p-value $< \alpha$
Two-tailed test: $Z_{cal} > Z_{\alpha/2}$

Example 7.1: An auditor claims that 10 percent of customers' ledger accounts carry mistakes in posting and balancing. A random sample of 600 was taken to test the accuracy of posting and balancing, and 45 mistakes were found. Are these sample results consistent with the auditor's claim? Use a 5 percent level of significance.
Solution: Let us take the null hypothesis that the claim of the auditor is valid, that is,

$$H_0 : p = 0.10 \quad \text{and} \quad H_1 : p \neq 0.10 \quad \text{(Two-tailed test)}$$

Given $\bar{p} = 45/600 = 0.075$, $n = 600$, and $\alpha = 5$ per cent. Thus using the z-test statistic

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{0.075 - 0.10}{\sqrt{\frac{0.10 \times 0.90}{600}}} = -\frac{0.025}{0.0122} = -2.049$$

Since z_{cal} (= − 2.049) is less than its critical (table) value z_α (= − 1.96) at $\alpha = 0.05$ level of significance, the null hypothesis, H0, is rejected. Hence, we conclude that the auditor's claim is not valid.

Example 7.2: A manufacturer claims that at least 95 percent of the equipment he supplied to a factory conformed to the specification. An examination of the sample of 200 pieces of equipment revealed that 18 were faulty. Test the manufacturer's claim.
Solution: Let us take the null hypothesis that at least 95 percent of the equipment supplied conformed to the specification, that is,

$$H_0 : p \geq 0.95 \quad \text{and} \quad H_1 : p < 0.95 \text{ (Left-tailed test)}$$

Given $\bar{p}$ = per cent of pieces conforming the specification = 1 – (18/100) = 0.91 n = 200 and level of significance α = 05. Thus using the z-test statistic,

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{0.91 - 0.95}{\sqrt{\frac{0.95 \times 0.05}{200}}} = -\frac{0.04}{0.015} = -2.67$$

Since mez_{cal} (= − 2.67) is less than its critical value z_α (= − 1.645) at α = 0.05 significance level, the null hypothesis, H_0, is rejected. Hence, we conclude that the proportion of equipment conforming to specifications is not 95%.

7.2 HYPOTHESIS TESTING FOR POPULATION MEAN WITH SMALL SAMPLES:

When the sample size is small (i.e., less than 30), the central limit theorem does not assure us that the sampling distribution of a statistic, such as the mean or proportion, is normal.

Consequently, when testing a hypothesis with small samples, we must assume that the samples come from a normally or approximately normally distributed population. Under these conditions, the sampling distribution of a sample statistic is normal, but the critical values of or$\bar{p}$ depend on whether or not the population standard deviation σ is known. When the value of the population standard deviation σ is unknown, its value is estimated by computing the standard deviation of sample s, and the standard error of the mean is calculated using the formula$\sigma_{\bar{x}} = s/\sqrt{n}$; when we do this, the resulting sampling distribution may not be normal even if sampling is done from a normally distributed population. In all such cases, the sampling distribution is the Student's *t-distribution.*

Sir William Gosset of Ireland, under his pen name 'Student,' developed a method for hypothesis testing popularly known as the 't-test' in the early 1900s. It is said that Guinness Brewery employed Gosset in Dublin, Ireland, which did not permit him to publish his research findings under his name, so he published his research findings in 1905 under the pen name 'Student.'

T-test: A hypothesis test for comparing two independent population means using the means of two small samples.

7.2.1 Uses of *t*-Distribution

There are various uses of *t*-distribution. A few of them are as follows:

1. Hypothesis testing for the population mean.
2. Hypothesis testing for the difference between two populations means using independent samples.
3. Hypothesis testing for the difference between two populations means using dependent samples.
4. Hypothesis testing for an observed correlation coefficient, including partial and rank correlations.
5. Hypothesis testing for an observed regression coefficient.

7.2.2 Hypothesis Testing for Single Population Mean

The test statistic for determining the difference between the sample mean $\bar{x}$ and population mean μ is given by

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}; \quad s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

Where s is an unbiased estimation of the unknown population standard deviation σ, this test statistic has a t-distribution with n –1 degrees of freedom.

Decision Rule: Rejected H_0 at the given degrees of freedom n–1 and level of significance when
One-tailed test: $t_{cal} > t_\alpha$ or p-value < α
Two-tailed test: $t_{cal} > t_{\alpha/2}$

Example 7.3: The average breaking strength of steel rods is specified to be 18.5 thousand kg. For this, a sample of 14 rods was tested. The mean and standard deviation obtained were 17.85 and 1.955, respectively. Test the significance of the deviation.

Solution: Let us take the null hypothesis that there is no significant deviation in the breaking strength of the rods, that is,

$H_0 : \mu = 18.5$ and $H_1 : \mu \neq 18.5$ (Two-tailed test)

Given, n = 14, $\bar{x}$ = 17.85, s = 1.955, df = n – 1 = 13, and α = 0.05 level of significance. The critical value *of* t at df = 13 and α/2 = 0.025 is $t_{\alpha/2}$ = 2.16.

Using the test statistic,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{17.85 - 18.5}{\frac{1.955}{\sqrt{14}}} = -\frac{0.65}{0.522} = -1.24$$

Since t_{cal} (= – 1.24) value is more than its critical value $t_{\alpha/2}$ = – 2.16 at α/2 = 0.025 and df = 13, the null hypothesis H_0 is accepted. Hence, we conclude that there is no significant deviation of the sample mean from the population mean.

Example 7.4: An automobile tire manufacturer claims that the average life of a particular grade of the tire is more than 20,000 km when used under normal conditions. A sample of 16 tires was tested, and a mean and standard deviation of 22,000 km and 5000 km were computed, respectively. Assuming the life of the tires in km to be approximately normally distributed, decide whether the manufacturer's claim is valid.

Solution: Let us take the null hypothesis that the manufacturer's claim is valid, that is,

$H_0 : \mu \geq 20{,}000$ and $H_1 : \mu < 20{,}000$ (Left-tailed test)

Given, n = 16, $\bar{x}$ = 22,000, s = 5000, df = 15 and α = 0.05 level of significance. The critical value *of* t at df = 15 and α = 0.05 is t_α = 1.753. Using the z-test statistic,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{22{,}000 - 20{,}000}{5000/\sqrt{16}} = \frac{2000}{1250} = 1.60$$

Since *t*he local (= 1.60) value is less than its critical value, t_α = 1.753, α = 0.05, and df = 15, the null hypothesis H_0 is accepted. Hence, we conclude that the manufacturer's claim is valid.

Example 7.5: A fertilizer mixing machine is set to give 12 kg of nitrate for every 100 kg of fertilizer. Ten bags of 100 kg each are examined. The percentage of nitrate obtained is 11, 14, 13, 12, 13, 12, 13, 14, 11, and 12. Is there reason to believe that the machine is defective?

Solution: Let us take the null hypothesis that the machine produces 12 kg of nitrate for every 100 kg of fertilizer and is not defective; that is,

$H_0 : \mu = 12$ and $H_1 : \mu \neq 12$ (Two-tailed test)

Given $n = 10$, $df = 9$, and $\alpha = 0.05$, critical value $t_{\alpha/2} = 2.262$ at $df = 9$ and $\alpha/2 = 0.025$. Assuming that the weight of nitrate in bags is usually distributed and its standard deviation is unknown, The sample mean $\bar{x}$ and standard deviation s values are calculated as shown in Table 7.1.

Table 7.1: Calculations of Sample Mean $\bar{x}$ and Standard Deviation s

Variable x	*deviation d= x-12*	d^2
11	-1	1
14	2	4
13	1	1
12	0	0
13	1	1
12	0	0
13	1	1
14	2	4
11	-1	1
12	0	0
		13

$$\bar{x} = \frac{\Sigma x}{n} = \frac{125}{10} = 12.5 \quad \text{and } s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}}$$

$$= \sqrt{\frac{13}{9} - \frac{(5)^2}{10(9)}} = 1.08$$

Using the test statistic, we have

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{12.5 - 12}{\frac{1.08}{\sqrt{10}}} = \frac{0.50}{0.341} = 1.466$$

Since t_{cal} (= 1.466) value is less than its critical value $t_{\alpha/2}$= 2.262, at $\alpha/2 = 0.025$ and df =9, the null hypothesis H_0 is accepted. Hence, we conclude that the manufacturer's claim is valid: the machine is not defective.

7.2.3 Hypothesis Testing for Difference of Two Population Means (Independent Samples)
To compare the mean values of two normally distributed populations, we draw independent random samples of sizes n_1 and n_2 from each population. If μ one and μ_2 are the mean values of two populations, then we aim to estimate the value of the difference $\mu_1 - \mu_2$ between the mean values of the two populations.

Since sample means $\bar{x}_1$ and $\bar{x}_2$ are the best point estimators to draw inferences regarding μ one and μ_2 respectively, therefore the difference between the sample means of the two independent simple random samples, $\bar{x}_1 - \bar{x}_2$, is the best point estimator of the difference $\mu_1 - \mu_2$.

The sampling distribution of $\bar{x}_1 - \bar{x}_2$ has the following properties:

- Expected value : $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$

This implies that the sample statistic $(\bar{x}_1 - \bar{x}_2)$ is an unbiased point estimator of $\mu_1 - \mu_2$.

- Variance : $\text{Var}(\bar{x}_1 - \bar{x}_2) = \text{Var}(\bar{x}_1) + \text{Var}(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

If the population standard deviations σ_1 and σ_2 are known, then the large sample interval estimation can also be used for the small sample case. However, if these are unknown, they are estimated by the sample standard deviations s_1 and s_2. It is needed if the sampling distribution is abnormal, even if sampling is done from two normal populations. This logic is the same as that for a single population case. Thus, *t*- t-distribution is used to develop a small sample interval estimate for $\mu_1 - \mu_2$.

Population Variances are Unknown, but Equal

If population variances σ_1^2 and σ_2^2 are unknown but equal, that is, both populations have the same shape and $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the standard error of the difference in two sample means $\bar{x}_1 - \bar{x}_2$ can be written as:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In such a case, we need not estimate σ12 and σ22 separately, and therefore, data from two samples can be combined to get a pooled, single estimate of σ^2. If we use the sample estimate s^2 for the population variance σ^2, then the pooled variance estimator of σ^2 is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

This single variance estimator, s^2, is a *weighted average* of the values of s_1^2 and s_2^2 in which weights are based on the degrees of freedom $n_1 - 1$ and $n_2 - 1$. Thus, the point estimate of $\sigma_{\bar{x}_1 - \bar{x}_2}$ when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is given by

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Since $s_1 = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1)^2}{(n_1 - 1)}}$ and $s_2 = \sqrt{\frac{\Sigma(x_2 - \bar{x}_2)^2}{(n_2 - 1)}}$,

Therefore, the pooled variance s^2 can also be calculated as

$$s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Following the same logic as discussed earlier, the *t*-test statistic is defined as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

The sampling distribution of this *t*-statistic is approximated by the *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Decision Rule Rejected H_0 at $df = n_1 + n_2 - 2$ and at specified level of significance α when
One-tailed test: $t_{cal} > t_\alpha$ or p-value < α
Two-tailed test: $t_{cal} > t_{\alpha/2}$

Confidence Interval: The confidence interval estimate of the difference between populations' means for small samples of size $n_1 < 30$ and/or $n_2 < 30$ with unknown σ_1 and σ_2 estimated by σ_1 and σ_2 is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\, s_{\bar{x}_1 - \bar{x}_2}$$

Where $t_{\alpha/2}$ is the critical value of t, the value of $t_{\alpha/2}$ depends on the t-distribution with $n_1 + n_2$ 2 degrees of freedom and confidence coefficient $1 - \alpha$.

Example 7.6: In a test given to two groups of students, the marks obtained are as follows:

First group	18	20	36	50	49	36	34	49	41
Second group	29	28	26	35	30	44	46		

Examine the significance of the difference between the arithmetic mean of the marks that the students of the above two groups secured.
Solution: Let us take the null hypothesis that there is no significant difference in the arithmetic mean of the marks secured by students of the two groups, that is,

$$H_0 : \mu_1 - \mu_2 = 0 \text{ or } \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2 \quad \text{(Two-tailed test)}$$

Since the sample size in both cases is small and the sample variances are unknown, the t-test statistic should be applied to test the null hypothesis.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Calculations of the sample mean $\bar{x}_1$, $\bar{x}_2$, and pooled sample standard deviations are shown in Table 7.2

Table 7.2: Calculation for $\bar{x}_1$, $\bar{x}_2$ and s

First Group x_1	$(x_1 - \bar{x}_1)$ $(x_1 - 37)$	$(x_1 - \bar{x}_1)^2$	***second group*** x_2	$(x_2 - \bar{x}_2)$ $(x_2 - 34)$	$(x_2 - \bar{x}_2)^2$
18	-19	361	29	-5	25
20	-17	389	28	-6	36
36	-1	1	26	-8	64
50	13	169	35	1	1
49	12	144	30	-4	16
36	-1	1	44	10	100
34	-3	9	46	12	144
49	12	144			
41	4	16			
333	**0**	**1234**	**238**	**0**	**386**

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{333}{9} = 37 \quad \text{and} \quad \bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{238}{7} = 34$$

$$s = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{1234 + 386}{9 + 7 - 2}} = 10.76$$

Substituting values in the *t*-test statistic, we get

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{37 - 34}{10.76}\sqrt{\frac{9 \times 7}{9 + 7}} = \frac{3}{10.46} \times 1.984 = 0.551$$

Degrees of freedom, $df = n_1 + n_2 - 2 = 9 + 7 - 2 = 14$
Since at $\alpha = 0.05$ and $df = 14$, the calculated value t_{cal} (= 0.551) is less than its critical value $t_{\alpha/2}$ = 2.14, the null hypothesis H_0 is accepted. Hence, we conclude that the mean marks obtained by the students of the two groups do not differ significantly.

7.2.4 Hypothesis Testing for Difference of Two Population Means (Paired t-test)

When two samples of the same size are paired so that each observation in one sample is associated with any particular observation in the second sample, the sampling procedure to collect the data and test the hypothesis is called *matched samples*. In such a case, the 'difference' between each data pair is first calculated. Then, these differences are treated as a single data set to consider whether there has been any significant change or whether the differences could have occurred by chance.

The matched sampling plan often leads to a more minor sampling error than the independent sampling plan because variation is eliminated as a source of sampling error in matched samples.

Let μ_d be the mean of the difference values for the population. Then, this mean value μ_d is compared to zero or some hypothesized value using the *t*-test for a single sample. The *t*-test statistic is used because the population's standard deviation of differences is unknown. Thus, the statistical inference about μ_d based on the average of the sample differences $\bar{d}$ would involve the *t*-distribution rather than the standard normal distribution. The *t*-test, also called *paired t-test*, becomes

$$t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}} / \sqrt{n}}$$

where
n = number of paired observations
$df = n - 1$, degrees of freedom
$\bar{d}$ = mean of the difference between paired (or related) observations
n = number of pairs of differences
s_d = sample standard deviation of the distribution of the difference between the paired (or related observations)

$$= \sqrt{\frac{\Sigma (d - \bar{d})^2}{n - 1}} = \sqrt{\frac{\Sigma d^2}{n - 1} - \frac{(\Sigma d)^2}{n (n - 1)}}$$

The null and alternative hypotheses are stated as follows:

$$H_0 : \mu_d = 0 \text{ or } c \text{ (Any hypothesized value)}$$
$$H_1 : \mu_d > 0 \text{ or } (\mu_d < 0) \text{ (One-tailed Test)}$$
$$\mu_d \neq 0 \text{ (Two-tailed Test)}$$

Decision rule: If the calculated value is less than its critical value, t_d, at a specified significance level and known degrees of freedom, the null hypothesis H_0 is accepted. Otherwise, H_0 is rejected.

Confidence interval: The confidence interval estimate of the difference between two population means is given by

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

Where $t_{\alpha/2}$ = critical value of the t-test statistic at $n - 1$ degrees of freedom and **a** significance level.
If the null hypothesis's claimed value, H0, lies within the confidence interval, then H0 is accepted; otherwise, it is rejected.

Example 7.7: The HRD manager wishes to see if trainees' abilities have changed after a specific training program. The trainees take an aptitude test before the start of the program and an equivalent one after completing it. The scores recorded are given below. Has any change taken place at a 5 percent significance level?

Trainee	A	B	C	D	E	F	G	H	I
Score before training	75	70	46	68	68	43	55	68	77
Score after training	70	77	57	60	79	64	55	77	76

Solution: Let us take the null hypothesis that no change has taken place after the training, that is,

$$H_0 : \mu_d = 0 \quad \text{and} \quad H_1 : \mu_d \neq 0 \quad \text{(Two-tailed test)}$$

The 'changes' are computed as shown in Table 7.3, and then a t-test is carried out on these differences as shown below.

Table 7.3: Calculations of 'Changes'

Trainee	***Before Training***	***After Traning***	***Difference in Scores, d***	d^2
A	75	70	5	25
B	70	77	7	49
C	46	57	-11	121
D	68	60	8	64
E	68	79	-11	121
F	43	64	-21	441
G	55	55	0	0
H	68	77	-9	81
I	77	76	1	1
			-45	903

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-45}{9} = -5 \text{ and}$$

$$s_d = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}} = \sqrt{\frac{903}{8} - \frac{(-45)^2}{9 \times 8}} = \sqrt{112.87 - 28.13} = 9.21$$

Applying the *t*-test statistic, we have

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{-5-0}{9.21/\sqrt{9}} = -\frac{5}{3.07} = -1.63$$

The null hypothesis is accepted since the calculated value $t_{cal} = -1.63$ is greater than its critical value, $t_{\alpha/2} = —2.31$, at $df = 8$ and $\alpha/2 = 0.025$. Hence, we conclude that trainees' abilities do not change after the training.

7.3 SELF-ASSESSMENT QUESTIONS:

1. A company manufacturing a specific type of breakfast cereal claims that 60 percent of all homemakers prefer that type to any other. A random sample of 300 homemakers contains 165 who like that type. Test the company's claim at a 5 percent level of significance.
2. An auditor claims that 10% of a company's invoices are incorrect. To test this claim, a random sample of 200 invoices is checked, and 24 are found wrong. At a 1 percent significance level, test whether the sample evidence supports the auditor's claim.
3. Ten oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 15.8 kg, and the standard deviation is 0.50 kg. Does the sample mean differ significantly from the intended weight of 16 kg?
4. Nine items in the sample had the following values: 45, 47, 50, 52, 48, 47, 49, 53, and 50. The mean is 49, and the sum of the squares of the deviation from the mean is 52. Can this sample be regarded as taken from the population with 47 as the mean? Also, 95 percent and 99 percent confidence limits of the population mean were obtained.
5. An IQ test was administered to 5 persons before and after training. The results are given below:
6.

Candidate	*I*	*II*	*III*	*IV*	*V*
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is any change in IQ level after the training program.

7.4 SUMMARY:

Hypothesis testing for a single sample proportion determines whether a population proportion matches a claimed value. The t-test is applied instead of the z-test for population means when dealing with small sample sizes. The Student's t-distribution is essential due to its flexibility with smaller datasets. It is beneficial when the population standard deviation is unknown.

Hypothesis testing can be performed for a single mean or to compare two means using independent or paired samples. These methods help make reliable inferences in various practical research contexts.

7.5 TECHNICAL TERMS:

- **Single Sample Proportion:** The proportion of a particular outcome in a single sample is tested to see if it matches a population claim.
- **Small Sample:** Generally refers to sample sizes less than 30, where normal approximation may not be appropriate.
- **Independent Samples:** Two unrelated or paired samples are used when comparing the means of two separate groups.
- **Paired Samples:** Two related samples, often the same group, measured before and after a treatment; analyzed with the paired t-test.
- **Critical Value:** The cutoff value that defines the boundary of the rejection region in hypothesis testing.
- **Significance Level (α):** The probability of rejecting the null hypothesis when it is true, commonly set at 0.05.

7.6 SUGGESTED READINGS:

1. Gupta, S. C., & Gupta, I. (2023). *Business Statistics* (18th Revised ed.). Himalaya Publishing House.
2. Bajpai, N. (2019). *Business Statistics* (2nd ed.). Pearson Education.
3. Sharma, J.K. (2020). *Business Statistics*. Pearson Education India.
4. Hooda, R. P. (2014). *Statistics for Business and Economics* (4th ed.). Macmillan Publishers India.

Dr. G. Malathi

LESSON- 8

F-DISTRIBUTION & CHI-SQUARE TEST

OBJECTIVES:

The purpose of studying this lesson is:

1. To understand the concept and properties of the F-distribution.
2. To learn to compare two population variances using the F-test.
3. To explore the structure and uses of the Chi-square test statistic.
4. To apply the chi-square test in various real-world statistical scenarios.
5. To conduct Chi-square tests of independence using contingency tables.
6. The Chi-square goodness-of-fit test will be used to assess how well-observed data match expected distributions.

STRUCTURE:

8.1 F-DISTRIBUTION:

We might need to compare population variances in several statistical applications. For instance, (i) variances in product quality resulting from two different production processes, (ii) variances in temperatures for two heating devices, (iii) variances in assembly times for two assembly methods, (iv) variance in the rate of return on investment of two types of stocks and so on, are few areas where comparison of variances is needed.

When independent random samples of size n_1 and n_2 are drawn from two normal populations, the ratio

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

follow F-distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$ degrees of freedom, where s_1^2 and s_2^2 are two sample variances and are given by

$$s_1^2 = \frac{\Sigma (x_1 - \bar{x}_1)^2}{n_1 - 1} \text{ and } s_2^2 = \frac{\Sigma (x_2 - \bar{x}_2)^2}{n_2 - 1}$$

F-test: A hypothesis test for comparing the variance of two independent populations with the help of variances of two small samples.
If two normal populations have equal variances, i.e., $\sigma_1^2 = \sigma_2^2$, then the ratio

$$F = \frac{s_1^2}{s_2^2} ; s_1 > s_2$$

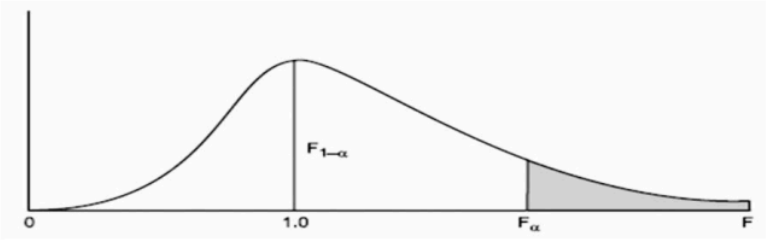
Probability distribution in repeated sampling is an F-distribution with n1 –1 degree of freedom for the numerator and n2 – 1 degree of freedom for the denominator. For computational purposes, a more significant sample variance is placed in the numerator so that the ratio is always equal to or greater than one.

Assumptions: A Few assumptions for the ratio s^2_1/s^2_2 to have an F-distribution are as follows:

1. Independent random samples are drawn from each of the two normal populations
2. The variability of the measurements in the two populations is the same and can be measured by a common variance σ^2, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$

The F-distribution, also called *variance ratio distribution*, is not symmetric, and the F values can never be negative. The shape of any F-distribution depends on the degrees of freedom of the numerator and denominator. A typical graph of an F-distribution is shown in Fig 8.1 for equal degrees of freedom for both numerator and denominator.

Figure 8.1 F-distribution for *n* Degrees of Freedom



8.1.1 Comparing Two Population Variances

How large or small must the ratio s^2_1/s^2_2 be for sufficient evidence to exist in the null?

Null hypothesis	Alternative hypothesis	
$H_0 : \sigma_1^2 = \sigma_2^2$	$H_1 : \sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$	(One-tailed Test)
$H_0 : \sigma_1^2 = \sigma_2^2$	$H_1 : \sigma_1^2 \neq \sigma_2^2$	(Two-tailed Test)

To conduct the test, random samples of size n_1 and n_2 are drawn from populations 1 and 2, respectively. The statistical test of the null hypothesis Ho uses the test statistic $F = s^2_1/s^2_2$, where s^2_1 and s^2_2 are the respective sample variances.

Decision rules: The criteria for acceptance or rejection of the null hypothesis H_0 are as follows:

1. Accept H_0 if the calculated value of the F-test statistic is less than its critical value $F\alpha$ (v1,v2), i.e., Fcal < $F\alpha$ for a one-tailed test.

The critical value of F_α is based on the degrees of freedom of the numerator, df1 = n1 – 1, and the degrees of freedom of the denominator, $df_2 = n_2 - 1$. These values can be obtained from F-Tables (See Appendix).

As mentioned earlier, the population with a more considerable variance is considered population 1 to ensure that a rejection of H_0 can occur only in the F-distribution curve's right (upper) tail. Even though half of the rejection region (the area α/2 to its left) will be in the lower tail of the distribution. It is never used because using the population with a more significant sample variance as population one always places the ratio s_1^2/s_2 in the right-tail direction.

2. $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_1 : \sigma_1^2 > \sigma_2^2$ (One-tailed test)

The null hypothesis is set up so that the rejection region is always in the upper tail of the distribution. This helps us consider the population with a larger variance in the alternative hypothesis.

Confidence Interval: An interval estimate of all possible values for a ratio σ^2_1/σ^2_2 of population variances is given by

$$\frac{s_1^2/s_2^2}{F_{(1-\alpha)}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2/s_2^2}{F_\alpha}$$

Where F values are based on an F-distribution with ($n_1 - 1$) and ($n_2 - 1$) degrees of freedom and (1 – α) confidence coefficient.

Example 8.1: Research was conducted to understand whether women have a more significant variation in attitude on political issues than men. The study used two independent samples of 31 men and 41 women. The sample variances calculated were 120 for women and 80 for men. Test whether the difference in attitude toward political issues is significant at the 5 percent level of significance.

Solution: Let us take the hypothesis that the difference is not significant, that is,

$H_0 : \sigma_w^2 = \sigma_m^2$ and $H_1 : \sigma_w^2 > \sigma_m$ (One-tailed test)

The F-test statistic is given by

$$F = \frac{s_1^2}{s_2^2} = \frac{120}{80} = 1.50$$

Since the variance for women is in the numerator, the one-tailed test will use the F-distribution with $df_1 = 41 - 1 = 40$ in the numerator and $df_2 = 31 - 1 = 30$ in the denominator.

The critical (table) value of $F_{\alpha=0.05} = 1.79$ at $df_1 = 40$ and $df_2 = 30$. The calculated value of F = 1.50 is less than its critical value of F = 1.79, so the null hypothesis is accepted. Hence, the research results do not support the belief that women have a more significant variation in attitudes on political issues than men.

Example 8.2: The following figures relate to the number of units of an item produced per shift by two workers, A and B, for several days

A	19	22	24	27	24	18	20	19	25		
B	26	37	40	35	30	30	40	26	30	35	45

Can it be inferred that worker A is more stable than worker B? Answer using the F-test at a 5 percent significance level.

Solution: Let us take the hypothesis that the two workers are equally stable, that is,

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ and } H_1 : \sigma_A^2 \neq \sigma_B^2 \quad \text{(One-tailed test)}$$

Table 8.1 shows the calculations for population variances σA2 and σB2 of the number of units produced by workers A and B, respectively.

Table 8.1: Calculation of σ_A^2 and σ_B^2

Worker A x_1	$(x_1 - \bar{x}_1)$ $(x_1 - 22)$	$(x_1 - 22)^2$	***Worker B*** x_2	$(x_2 - \bar{x}_2)$ $(x_2 - 34)$	$(x_2 - \bar{x}_2)^2$
19	-3	9	26	-8	64
22	0	0	37	3	9
24	2	4	40	6	36
27	5	25	35	1	1
24	2	4	30	-4	16
18	-4	16	30	-4	16
20	-2	4	40	6	36
19	-3	9	26	-8	64
25	3	9	30	-4	16
			35	1	1
			45	11	121
198	**0**	**80**	**374**	**0**	**380**

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{198}{9} = 22; \qquad \bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{374}{11} = 34$$

$$s_A^2 = \frac{\sum (x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{80}{9-1} = 10; \qquad s_B^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{380}{11-1} = 38$$

Applying the F-test statistic, we have

$$F = \frac{s_B^2}{s_A^2} = \frac{38}{10} = 3.8 \text{ (because } s_B^2 > s_A^2)$$

The critical value $F_{0\ 05(10\ 8)} = 3.35$ at $\alpha = 5$ percent significance level and degrees of freedom $df_A = 8$, $df_B = 10$. The null hypothesis is rejected since the calculated value of F is more than its critical value. Hence, we conclude that Worker A is more stable than Worker B because $s_A^2 < s_B^2$

8.2 THE CHI-SQUARE TEST-STATISTIC:

Like *t* and F distributions, a *χ2-distribution* is also a function of its degrees of freedom. This distribution is skewed to the right; the random variable can never take a negative value. Theoretically, its range is from 0 to ∞ as shown below. Values of x^2 that divide the curve with a proportion of the area equivalent to α (level of significance) in the right tail are given in the Appendix. The x^2-test statistic is given by

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

Where

O = an observed frequency in a particular category

E = an expected frequency for a particular category

Decision Rule: The calculated value of x^2 is compared with its critical value at a particular significance level and degrees of freedom. If $x^2_{cal} > x_{critical}^2$, then the null hypothesis is rejected in favor of the alternative hypothesis, and it is concluded that the difference between the two sets of frequencies is significant.

8.2.1 Applications of X^2 Test

A few essential applications of x^2 test discussed in this chapter are as follows:

- Test of independence
- Test of goodness-of-fit
- Yate's correction for continuity
- Test for population variance
- Test for homogeneity

8.2.2 Contingency Table Analysis: Chi-Square Test of Independence

The x^2 test of independence is used to analyze the frequencies of two qualitative variables or attributes with multiple categories to determine whether the two variables are independent. The chi-square test of independence can be used to analyze any level of measurement, but it is instrumental in analyzing nominal data. For example,

- Whether voters can be classified by gender is independent of their political affiliation
- Whether university students classified by gender are independent of the courses of study
- Whether wage-earners classified by education level are independent of income
- The type of soft drink a consumer prefers is independent of the consumer's age.
- Whether absenteeism is independent of job classification
- Whether an item is acceptable is independent of the shifts in which it was manufactured.

When observations are classified according to two qualitative variables or attributes and arranged in a table, the display is called a contingency table, as shown in Table 8.2. The test of independence uses the contingency table format and is also referred to as a *Contingency Table Analysis (or Test)*.

Contingency table: A cross-table for displaying the frequencies of all possible groups of two variables.

Table 8.2: Contingency Table

Variable B	*Variable A*				Total
	A_1	A_2	A_3 ……	A_C	
B_1	O_{11}	O_{12}	O_{13}	O_{1C}	R_1
B_2	O_{21}	O_{22}	O_{23}.....	O_{2c}	R_2
.	.				.
.					.
.					.
B_r	O_{r1}	O_{r2}	O_{r3}	O_{rC}	. R_r
Total	C_1	C_2	C_3 ……..	C_c	N

It may be noted that variables A and B have been classified into mutually exclusive categories. The value O_{ij} is the observed frequency for the cell in row *i* and column *j*. The row and column totals are the sums of the frequencies. The row and column totals are added to get a total of n, representing the sample size.

The *expected frequency*, E_{ij}, corresponding to an observed frequency O_{ij} in row *i* and column *j* under the assumption of independence, is based on the multiplicative rule of probability. If two events, A_i and B_j, are independent, then the likelihood of their joint occurrence is equal to the product of their probabilities. Thus, the expected frequencies in each cell of the contingency table are calculated as follows:

$$E_{ij} = \frac{\text{Row } i \text{ total}}{\text{Sample size}} \times \frac{\text{Column } j \text{ total}}{\text{Sample size}} \times \text{Grand total}$$

$$= \frac{R_i}{N} \times \frac{C_j}{N} \times N = \frac{R_i \times C_j}{N}$$

The analysis of a two-way contingency table helps to answer the question of whether the two variables are unrelated or independent of each other. Consequently, *the null hypothesis for a chi-square test of independence is that the two variables are independent.* If the null hypothesis H0 is rejected, then the two variables are not independent but are related. Hence, the x^2-test statistic measures how much the observed frequencies differ from the expected frequencies when the variables are independent.

The Procedure The Procedure to test the association between two independent variables, where the sample data is presented in the form of a contingency table with *r* rows and *c* columns, is summarized as follows:

State the null and alternative hypotheses

Step :

H_0: No relationship or association exists between two variables; that is, they are independent

H_1: A relationship exists; that is, they are related

Step 2: Select a random sample, record the observed frequencies (O values) in each contingency table cell, and calculate the row, column, and grand totals.

Step 3: Calculate the expected frequencies (*E*-values) for each cell:

$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Step 4: Compute the value of the test statistic

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

Step 5: Calculate the degrees of freedom. The formula gives the degrees of freedom for the chi-square test of independence.

df = (Number of rows – 1) (Number of columns – 1) = (*r* – 1) (*c* – 1)

Step 6: Using a level of significance α and *df*, find the critical (table) value of x_α^2 (see Appendix). This value of x^2_α corresponds to an area in the distribution's right tail.

Step 7: Compare the calculated and table values of x^2. Decide whether the variables are independent or not using the decision rule:

- Accept H_0 if x^2_{cal} is less than its table value $x^2_{\alpha,\,(r-1)\,(c-1)}$
- Otherwise, reject H_0

Example 8.2: Two hundred randomly selected adults were asked whether TV shows are primarily entertaining, educational, or a waste of time (only one answer could be chosen). The respondents were categorized by gender. Their responses are given in the following table:

Gender	*Opinion*			
	Entertaining	*Educational*	*Waste of Time*	*Total*
Male	52	28	30	110
Female	28	12	50	90
Total	80	40	80	200

Is this evidence convincing that there is a relationship between gender and opinion on the population's interest?

Solution: Let us assume the null hypothesis that the opinion of adults is independent of gender. The contingency table is of size 2×3; the degrees of freedom would be (2–1) (3–1) = 2; we will have to calculate only two expected frequencies. The other four can be automatically determined as shown below:

$$E_{11} = \frac{\text{Row 1 total} \times \text{Column 1 total}}{\text{Grand total}} = \frac{110 \times 80}{200} = 44$$

$$E_{12} = \frac{\text{Row 1 total} \times \text{Column 2 total}}{\text{Grand total}} = \frac{110 \times 40}{200} = 22$$

$$E_{13} = 110 - (44 + 22) = 44$$

$$E_{21} = 80 - E_{11} = 80 - 44 = 36$$

$$E_{22} = 40 - E_{12} = 40 - 22 = 18$$

$$E_{23} = 80 - E_{13} = 80 - 44 = 36$$

The contingency table of expected frequencies is as follows:

Gender	*Opinion*			
	Entertaining	*Educational*	*Waste of Time*	*Total*
Male	44	22	44	110
Female	36	18	36	90
Total	80	40	80	200

Arranging the observed and expected frequencies in the following table to calculate the value of the x^2-test statistic:

Observed (O)	***Expected (E)***	***O − E***	***(O − E)²***	***(O − E)² / E***
52	44	8	64	1.454
28	22	6	36	1.636
30	44	14	196	4.455
28	36	-8	64	1.777
12	18	-6	36	2.000
50	36	14	196	5.444
				16.766

The critical (or table) value of $x^2 = 5.99$ at $\alpha = 0.05$ and $df = 2$. The null hypothesis is rejected since the calculated value of x2 = 16.766 is more than its critical value. Hence, we conclude that the opinion of adults is not independent of gender.

Example 8.3: A company is interested in determining whether an association exists between the commuting time of their employees and the level of stress-related problems observed on the job. A study of 116 assembly-line workers reveals the following:

Counting Time	*Stress*			
	High	*Moderate*	*Low*	*Total*
Under 20 min	9	5	18	32
20-50 min	17	8	28	53
Over 50 min	18	6	7	31
Total	44	19	53	116

At $\alpha = 0.01$ level of significance, is there any evidence of a significant relationship between commuting time and stress?

Solution: Let us assume the null hypothesis that stress on the job is independent of commuting time.

The contingency table is of size 3×3, and the degrees of freedom would be (3 – 1) (3 – 1) = 4; that is, we will have to calculate only four expected frequencies, and the others can be calculated automatically as shown below:

$$E_{11} = \frac{32 \times 44}{116} = 12.14 \quad E_{12} = \frac{32 \times 19}{116} = 5.24 \quad E_{13} = 14.62$$

$$E_{21} = \frac{53 \times 44}{116} = 20.10 \quad E_{22} = \frac{53 \times 19}{116} = 8.68 \quad E_{23} = 24.22$$

$$E_{31} = \frac{31 \times 44}{116} = 11.75 \quad E_{32} = \frac{31 \times 19}{116} = 5.08 \quad E_{33} = 14.17$$

Arranging the observed and expected frequencies in the following table to calculate the value of the x^2-test statistic:

Observed (O)	***Expected (E)***	***O − E***	***(O − E)²***	***(O − E)² / E***
9	12.14	−3.14	9.85	0.811
5	5.24	−0.24	0.05	0.009
18	14.62	3.38	11.42	0.781
17	20.10	−3.10	9.61	0.478
8	8.68	−0.68	0.45	0.052
28	24.22	3.78	14.28	0.589
18	11.75	6.25	39.06	3.324
6	5.08	0.92	0.84	0.165
7	14.17	−7.17	51.40	3.627
				9.836

The critical value of $x^2 = 13.30$ at $\alpha = 0.01$ and $df = 4$. Since the calculated value of $x^2 = 9.836$ is less than its critical value, the null hypothesis H_0 is accepted. Hence, we conclude that stress on the job is independent of commuting time.

8.2.3 Chi-Square Test for Goodness-of-Fit

On several occasions, a decision-maker must understand whether an actual sample distribution matches or coincides with a known theoretical probability distribution, such as binomial, Poisson, normal, etc. *The x^2 test for* **goodness-of-fit** *is a statistical test of how well-given data supports an assumption about the distribution of a population or random variable of interest.*

The test determines how well an assumed distribution fits the provided data. To apply this test, a particular theoretical distribution is first hypothesized for a given population, and then the test is carried out to determine whether or not the sample data could have come from the population of interest with the hypothesized theoretical distribution. The observed frequencies or values come from the sample, and the expected ones come from the theoretical hypothesized probability distribution. The goodness-of-fit test now focuses on the differences between the observed and predicted values. Significant differences between the two distributions cause doubt about the assumption that the hypothesized theoretical distribution is correct. On the other hand, slight differences between the two distributions may be assumed to be resulting from sampling error.

Goodness-of-fit: A statistical test conducted to determine how closely the observed frequencies fit those predicted by a hypothesized probability distribution for the population.

The Procedure The general steps to conduct a goodness-of-fit test for any hypothesized population distribution are summarized as follows:

Step 1: State the null and alternative hypotheses

H_0: No difference between the observed and expected sets of frequencies.

H_1: There is a difference

Step 2: Select a random sample and record each category's observed frequencies (O values).

Step 3: Calculate expected frequencies (E values) in each category by multiplying the category probability by the sample size.

Step 4: Compute the value of the test statistic

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

Step 5: Using a significance level α and $df = n - 1$, provided that the number of expected frequencies is five or more for all categories, find the critical (table) value of x^2 (See Appendix).

Step 6: Compare the calculated and table value of x^2, and use the following decision rule:

- Accept H_0 if x_{cal}^2 is less than its critical value $x^2_{\alpha, n-1}$
- Otherwise, reject H_0

Example 8.4: A Personnel Manager is interested in determining whether absenteeism is more significant on one day of the week than another. His records for the past year show the following sample distribution:

Day of the Week	Monday	Tuesday	Wednesday	Thursday	Friday
No. of Absentees	66	56	54	48	75

Test whether the absence is uniformly distributed over the week

Solution: Let us assume the null hypothesis that the absence is uniformly distributed over the week.

The number of absentees during a week is 300, and if absenteeism is equally probable on all days, then we should expect 300/5 = 60 absentees on each day of the week. Now arrange the data as follows:

Category	O	E	O − E	$(O - E)^2$	$(O - E)^2 / E$
Monday	66	60	6	36	0.60
Tuesday	57	60	−3	9	0.15
Wednesday	54	60	−6	36	0.60
Thursday	48	60	−12	144	2.40
Friday	75	60	15	225	3.75
Total					7.50

The critical value of $x^2 = 9.49$ at $\alpha = 0.05$ and $df = 5 - 1 = 4$. The null hypothesis is accepted since the calculated value $x^2\text{cal} = 7.50$ is less than its critical value.

8.3 SUMMARY:

The F-distribution is used primarily to compare two population variances through the F-test. It helps determine whether the variability in the two groups is significantly different. The Chi-square test is a non-parametric method used to evaluate categorical data. Applications include testing independence between variables using contingency tables and checking how well-observed frequencies fit expected distributions with the goodness-of-fit test. These tools are essential in analyzing relationships and patterns in categorical data. They support decisions in genetics, market research, and social sciences.

8.4 TECHNICAL TERMS:

- **F-Distribution:** A probability distribution used to compare two variances; positively skewed and depends on degrees of freedom.
- **F-Test:** A statistical test using the F-distribution to assess if two population variances are significantly different.
- **Variance:** A measure of how data points differ from the mean; used to assess data spread.
- **Chi-Square Test-Statistic (χ^2):** A value used to assess differences between observed and expected frequencies in categorical data.
- **Applications of Chi-Square:** Includes tests for independence, goodness-of-fit, and homogeneity.
- **Contingency Table**: A table that displays frequencies for combinations of two categorical variables used in the chi-square test of independence.
- **Chi-Square Test of Independence:** Determines whether two categorical variables are statistically independent.
- **Goodness-of-Fit Test:** Evaluates how well-observed data fit a theoretical distribution.
- **Degrees of Freedom:** A parameter used in statistical distributions affects the shape of the F and chi-square distributions.
- **Expected Frequency:** The theoretically predicted frequency in a category under a specific hypothesis.
- **Observed Frequency:** The actual frequency counted in a sample for a given category.

8.5 SELF-ASSESSMENT QUESTIONS:

1. The mean diameter of a steel pipe produced by processes A and B is practically the same, but the standard deviations may differ. For a sample of 22 pipes produced by A, the standard deviation is 2.9 m, while for a sample of 16 pipes produced by B, the standard deviation is 3.8 m. Test whether the pipes made by process A have the same variability as those produced by process B.

2. Two random samples drawn from the normal population are:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Sample 1	20	16	26	27	23	22	18	24	25	19		
Sample 2	27	33	42	35	32	34	38	28	41	43	30	37

Obtain estimates of the population variances and test whether the two populations have the same variance.

3. In an anti-malaria campaign in a specific area, quinine was administered to 812 persons out of a total population of 3248. The number of fever cases reported is shown below:

Treatment	*Fever*	*No Fever*	*Total*
Quinine	20	792	812
No quinine	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of quinine in checking malaria.

4. Based on information from 1000 randomly selected fields about the tenancy status of the cultivation of these fields and the use of fertilizers collected in an agro-economic survey, the following classifications were noted:

	Owned	*Rented*	*Total*
Using fertilizers	416	184	600
Not using fertilizers	64	336	400
Total	480	520	1000

Would you conclude that owner cultivators are more inclined to use fertilizers at the $\alpha = 0.05$ level of significance? Carry out the chi-square test as per the testing procedures.

5. The demand for a particular spare part in a factory was found to vary from day to day. In a sample study, the following information was obtained:

Day	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
Number of Parts Demanded	1124	1125	1110	1120	1126	1115

6. Describe how the chi-square procedure determines the expected frequencies.

7. What is the x^2-test? Under what conditions is it applicable? Point out its role in business decision-making.

8. Describe the X^2-test of significance and state the various uses.

8.6 SUGGESTED READINGS:

1. Gupta, S. C., & Gupta, I. (2023). *Business Statistics* (18th Revised ed.). Himalaya Publishing House.
2. Bajpai, N. (2019). *Business Statistics* (2nd ed.). Pearson Education.
3. Sharma, J.K. (2020). *Business Statistics*. Pearson Education India.
4. Hooda, R. P. (2014). *Statistics for Business and Economics* (4th ed.). Macmillan Publishers India.

Dr. G. Malathi

LESSON- 9

CORRELATION

OBJECTIVES:

The purpose of studying this lesson is:

- To understand the concept and significance of the coefficient of correlation.
- To distinguish between different types of correlations (positive, negative, linear, and nonlinear).
- To construct and interpret scatter diagrams to observe data relationships.
- To calculate Karl Pearson's correlation coefficient and interpret its value.
- To understand the concept of the coefficient of determination and its role in explaining variability.

STRUCTURE:

9.1 INTRODUCTION TO CORRELATION:

The statistical methods introduced in this lesson are designed to simultaneously analyze data involving two quantitative variables. Examining a single variable in isolation is not enough in many real-world scenarios. Instead, understanding how variables relate can provide deeper insights into patterns, trends, and potential causal relationships within the data. Identifying such associations allows us to draw meaningful inferences, supporting more informed, data-driven decision-making.

This brings us to the concept of **correlation**, a fundamental statistical tool used to measure the strength and direction of a linear relationship between two quantitative variables. Understanding these relationships can be critical in a variety of fields. For example:

- The economic relationship between consumer income and spending habits can guide fiscal policy decisions.
- Studying the association between physical activity levels and blood pressure in healthcare may inform public health recommendations.

- Exploring the correlation between study time and academic performance in education can help design more effective learning strategies.
- Analyzing the relationship between advertising expenditure and sales revenue in business can support strategic marketing decisions.

Through correlation analysis, we gain valuable tools to interpret real-world data more effectively and apply this understanding to practical problems across disciplines.
Correlation is a statistical measure that describes the strength and direction of a linear relationship between two quantitative variables. It indicates how changes in one variable are associated with changes in another. The value of the correlation coefficient typically ranges from 1 to +1:

- A value close to +1 indicates a strong positive linear relationship (as one variable increases, the other also increases),
- A value close to 1 indicates a strong negative linear relationship (as one variable increases, the other decreases),
- A value around 0 suggests little to no linear relationship between the variables.

9.1.1 Coefficient of correlation: A statistical measure of the degree of association between two variables

- When a pair of values of two variables are plotted on a graph, the **strength** of the relationship is determined by the closeness of the points to a straight line. A straight line is used as the frame of reference for evaluating the relationship.
- The ***direction*** is determined by whether one variable generally increases or decreases when the other variable increases.

The importance of examining the statistical relationship between two or more variables can be divided into the following questions, and accordingly requires the statistical methods to answer these questions:

1. Is there an association between two or more variables? If yes, what is the form and degree of that relationship?
2. Is the relationship strong or significant enough to help arrive at a desirable conclusion?

For correlation analysis, the data on values of two variables must come from sampling in pairs, one for each of the two variables. The pairing relationship should represent some time, place, or condition.

9.2 TYPES OF CORRELATIONS:

There are three broad types of correlations:

1. Positive and negative,
2. Linear and nonlinear,
3. Simple, partial, and multiple.

This lesson will discuss simple linear positive or negative correlation analysis.

Positive and Negative Correlation
A positive (or direct) correlation refers to the same direction of change in the values of variables. In other words, if the values of variables are varying (i.e., increasing or decreasing) in the same direction, then such a correlation is referred to as a **positive correlation**.
A **negative (or inverse) correlation** refers to the change in the values of variables in opposite directions.

The following examples illustrate the concept of positive and negative correlation.

Positive Correlation
Increasing → x: 5 8 10 15 17
Increasing → y: 10 12 16 18 20
Decreasing → x: 17 15 10 8 5
Decreasing → y: 20 18 16 12 10

Negative Correlation
Increasing → x: 5 8 10 15 17
Decreasing → y: 20 18 16 12 10
Decreasing → x: 17 15 12 10 6
Increasing → y: 2 7 9 13 14
The change (increasing or decreasing) in values of both variables is not proportional or fixed.

Linear and Nonlinear Correlation
A linear correlation implies a constant change in one variable's values with respect to a change in the corresponding values of another variable. In other words, a correlation is referred to as *linear correlation* when variations in the values of two variables have a constant ratio. The following example illustrates a linear correlation between the x and y variables.

X	*10*	*20*	*30*	*40*	*50*
Y	*40*	*60*	*80*	*100*	*120*

When these pairs of values of x and y are plotted on graph paper, the line joining these points is straight.

A nonlinear (or curvilinear) correlation implies an absolute change in one variable's values concerning changes in the values of another variable. In other words, a correlation is referred to as a *nonlinear correlation* when the amount of change in the values of one variable does not bear a constant ratio to the amount of change in the corresponding values of another variable. The following example illustrates a nonlinear correlation between x and y variables.

X	*8*	*9*	*9*	*10*	*10*	*28*	*29*	*30*
Y	*80*	*130*	*170*	*150*	*230*	*560*	*460*	*600*

When these x and y values are plotted on graph paper, the line joining these points is not straight but curvilinear.

Simple, Partial, and Multiple Correlation
The distinction between simple, partial, and multiple correlations is based on the number of variables involved in the correlation analysis.
As the number of hours a student studies, their exam scores tend to grow.
This indicates a **positive correlation** between study time and exam performance.
It suggests that the two variables move in the same direction.

Two variables are chosen to study the correlation between them in partial correlation, but the effect of other influencing variables is kept constant. For example, (i) the yield of a crop is influenced by the amount of fertilizer applied, rainfall, quality of seed, type of soil, and pesticides, (ii) sales revenue from a product is influenced by the level of advertising expenditure, quality of the product, price, competitors, distribution, and so on. In such cases,

an attempt to measure the correlation between yield and seed quality, assuming that the average values of other factors exist, becomes a problem of partial correlation.
The relationship between more than three variables is considered simultaneously for study in multiple correlations. For example, the employer-employee relationship in any organization may be examined concerning training and development facilities, medical, housing, and education facilities for children, salary structure, grievance handling system, and so on.

9.3 METHODS OF CORRELATION ANALYSIS:

The correlation between two ratio-scaled (numeric) variables is represented by the letter *r*, which only takes on values between -1 and +1. Sometimes, this measure is called the '**Pearson product-moment correlation**' or the **correlation coefficient**. The correlation coefficient is scale-free; therefore, its interpretation is independent of the units of measurement of two variables, say, *x* and *y*.

In this chapter, the following methods of finding the correlation coefficient between two variables, *x* and *y*, are discussed:

1. Scatter Diagram Method
2. Karl Pearson's Coefficient of Correlation method
3. Spearman's Rank Correlation method
4. Method of Least Squares

9.3.1 Scatter Diagram Method

The **scatter diagram** is a quick, at-a-glance method of determining an apparent relationship between two variables, if any. A scatter diagram (or a graph) can be obtained on a graph paper by plotting observed (or known) pairs of variables x and y values, taking the independent variable values on the *x*-axis and the dependent variable values on the *y*-axis.

Scatter diagram: A graph of pairs of values of two variables that are plotted to indicate a visual display of the pattern of their relationship.
It is common to try to draw a straight line through data points so that an equal number of points lie on either side of the line. This straight line defines the relationship between two variables, x and y, described by the data points.

In a scatter diagram, the horizontal and vertical axes are scaled in units corresponding to the variables *x* and *y*, respectively. Figure 9.1 shows examples of different types of relationships based on pairs of x and y values in a sample dataset. The pattern of data points in the diagram indicates that the variables are related. If the variables are related, the dotted line in each diagram describes the relationship between the two variables.

The patterns depicted in Fig. 9.1(a) and (b) represent linear relationships since straight lines describe the patterns. The pattern in Fig. 9.1(a) shows a *positive* relationship since the value of *y* tends to increase as the value of *x* increases. In contrast, the pattern in Fig. 9.1(b) shows a *negative* relationship since the value of *y* tends to decrease as the value of *x* increases.

The pattern depicted in Fig. 9.1(c) illustrates a very low or no relationship between x and y values. In contrast, Fig. 9.1(d) represents a curvilinear relationship since it is described by a curve rather than a straight line. Fig. 9.1(e) illustrates a positive linear relationship with a widely scattered pattern of points. The wider scattering indicates a lower degree of association between the two variables, x, and y, than in Fig. 9.1(a).

Interpretation of Correlation Coefficients: While interpreting the correlation coefficient *r*, the following points should be taken into account:

1. A low r-value does not indicate that the variables are unrelated but suggests that a straight line poorly describes the relationship. A nonlinear relationship may also exist.
2. A correlation does not imply a *cause-and-effect* relationship; it is merely an observed association.

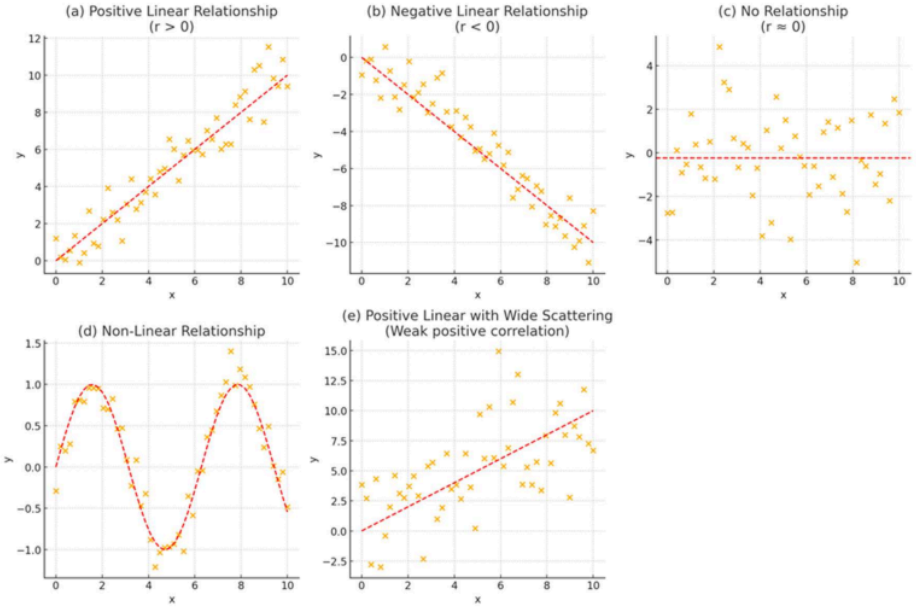


Figure 9.1 Examples of Correlation Coefficient

Example 9.1: Given the following data:

Student	1	2	3	4	5	6	7	8	9	10
Aptitude Score	400	675	475	350	425	600	550	325	675	450
Grade Point Average	1.8	3.8	2.8	1.7	2.8	3.1	2.6	1.9	3.2	2.3

1. Draw this data on graph paper.
2. Is there any correlation between aptitude score and grade point average? If yes, what is your opinion?

Solution: By taking an appropriate scale on the *x* and *y* axes, the pair of observations are plotted on graph paper, as shown in Fig. 13.3. The scatter diagram in Fig. 13.3, with a straight line representing the relationship between x and y fitted through it, is shown in Fig. 13 .3.

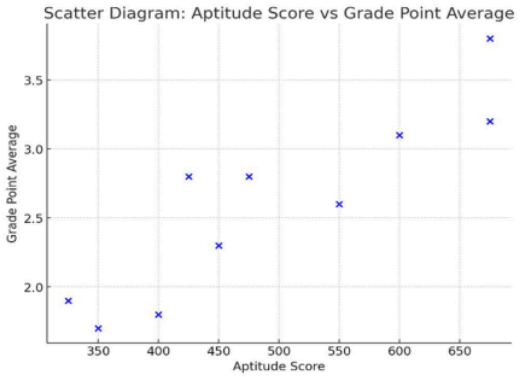


Figure 9.2 Scatter Diagram

Interpretation: From the scatter diagram, there appears to be a high degree of association between the two variable values. It is because the data points are very close to a straight line passing through the points. This pattern of dotted points also indicates a high degree of linear positive correlation.

9.3.2 Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient measures quantitatively the extent to which two variables, x and y, correlate. For a set of n pairs of values of x and y, Pearson's correlation coefficient r is given by:

Pearson Correlation Coefficient Formula

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{var } x}\sqrt{\text{var } y}} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

where $$\text{Cov}(x, y) = \frac{1}{n}\sum (x - \bar{x})(y - \bar{y})$$

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \leftarrow \text{standard deviation of sample data on variable } x$$

$$\sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} \leftarrow \text{standard deviation of sample data on variable } y$$

Substituting the mathematical formula for Cov (x, y) and σ_x and σ_y, we have

$$r = \frac{\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\frac{1}{n}\sum (x - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum (y - \bar{y})^2}} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}}$$

Step Deviation Method for Ungrouped Data: When actual mean values are in fractions, the calculation of Pearson's correlation coefficient can be simplified by taking deviations of x and y values from their assumed means, A and B, respectively. That is, $d_x = x - A$ and $d_y = y - B$, where A and B are assumed means of x and y values. The formula (13-1) becomes

$$r = \frac{n\sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{n\sum d_x^2 - (\sum d_x)^2} \cdot \sqrt{n\sum d_y^2 - (\sum d_y)^2}}$$

Step Deviation Method for Grouped Data When data on x and y values are classified or grouped into a frequency distribution, the formula (13-2) is modified as follows:

$$r = \frac{n\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{n\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

Probable Error and Standard Error of the Coefficient of Correlation

The probable error (PE) of the coefficient of correlation indicates the extent to which its value depends on the random sampling condition. If r is the calculated value of the correlation coefficient in a sample of n pairs of observations, then the standard error SE_r of the correlation coefficient r is given by

$$SE_r = \frac{1-r^2}{\sqrt{n}}$$

The expression calculates the probable error of the coefficient of correlation:

$$PE_r = 0.6745\, SE_r = 0.6745\, \frac{1-r^2}{\sqrt{n}}$$

Thus, with the help of PE_r, we can determine the range within which the population coefficient of correlation is expected to fall using the following formula:

$$\rho = r \pm PE_r$$

Where ρ (rho) represents the population coefficient of correlation.

Remarks

1. If $r < PE_r$, then the r value is insignificant; there is no relationship between the two variables of interest.
2. If $r > 6PE_r$, then the r value is significant; a relationship exists between the two variables.

Illustration: If $r = 0.8$ and $n = 25$, then PE_r is

$$PE_r = 0.6745\, \frac{1-(0.8)^2}{\sqrt{25}} = 0.6745\, \frac{0.36}{5} = 0.048$$

Thus, the limits within which the population correlation coefficient (ρ_r) should fall are

$$r \pm PE_r = 0.8 \pm 0.048 \quad \text{or} \quad 0.752 \le \rho_r \le 0.848$$

9.3.2.1 The Coefficient of Determination

The squared value of the correlation coefficient r is called **the coefficient of determination, denoted as r^2; It** always has a value between 0 and 1. By squaring the correlation coefficient, we retain information about the strength of the relationship, but we lose information about the direction. *This measure represents the proportion (or percentage) of the total variability of the dependent variable, y, that is accounted for or explained by the independent variable, x.* The

proportion (or percentage) of variation in *y* that *x* can explain determines the extent or strength of association between the two variables *x* and *y*.

- The coefficient of correlation r has been grossly overrated and used too much. Its square, the coefficient of determination r^2, is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before concluding about the extent of the linear relationship between two correlated variables.

Coefficient of determination: A statistical measure of the proportion of the variation in the dependent variable explained by the independent variable.

Interpretation of Coefficient of Determination: The Coefficient of determination is preferred for interpreting the strength of association between two variables because it is easier to analyze a percentage. Fig. 9.3 illustrates the meaning of the coefficient of determination:

- If $r^2 = 0$, then *no variation* in *y* can be explained by the variable *x*.
- If $r^2 = 1$, then the values of *y* are *entirely explained* by *x*. There is a *perfect association* between *x* and *y*.
- If $0 \leq r^2 \leq 1$, the degree of explained variation in *y* due to *variation in values of x* depends on the value of r^2. A value of r2 closer to 0 shows a low proportion of variation in *y*, as explained by *x*. On the other hand, the value of r^2 closer to 1 shows that variable *x* can predict the actual value of the variable *y*

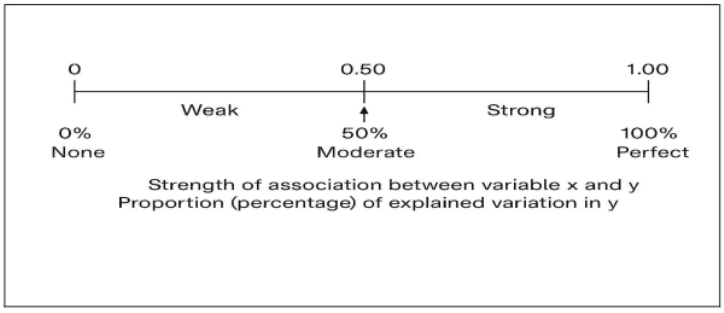


Figure Interpretation of Coefficient of Determination

Mathematically, the coefficient of determination is given by

$$r^2 = 1 - \frac{\text{Explained variability in } y}{\text{Total variability in } y}$$

$$= 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{n\Sigma y^2 - a\Sigma y - b\Sigma xy}{n\Sigma y^2 - (\bar{y})^2}$$

Where is the estimated value of *y* for given values of *x*? One minus the ratio between these two variations is called the *coefficient of determination*.

For example, let the correlation between variable x (height) and variable y (weight) be r = 0.70. Now, the coefficient of determination r2 = 0.49, or 49 percent, implies that only 49 percent of the variation in variable y (weight) can be accounted for in terms of variable x (height). The remaining 51 percent of the variability may be due to other factors, such as the tendency to eat fatty foods.

Even a relatively high correlation coefficient $r = 0.70$ accounts for less than 50 percent of the variability. In this context, it is essential to know that 'variability' refers to how values of variable y are scattered around its mean value. As in the above example, some people will be heavy, some average, some light. So, we can account for 49 percent of the total weight variability (y) in height (x) if r=0.70. The greater the correlation coefficient, the greater the coefficient of determination and the variability in the dependent variable can be accounted for in terms of the independent variable.

Example 9.2: The following table gives indices of industrial production and the number of registered unemployed people (in lakhs). Calculate the value of the correlation coefficient.

Year	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**
Index of Production	100	102	104	107	105	112	103	99
Number Unemployed	15	12	13	11	12	12	19	26

Solution: Calculations of Karl Pearson's correlation coefficient are shown in the table below:

Year	***Production (x)***	***dx = (x - $\bar{x}$)***	***d_x^2***	***Unemployed (y)***	***dy = (y - $\bar{y}$)***	***d_y^2***	***$d_x \cdot d_y$***
1991	100	−4	16	15	0	0	0
1992	102	−2	4	12	−3	9	+6
1993	104	0	0	13	−2	4	0
1994	107	+3	9	11	−4	16	−12
1995	105	+1	1	13	−2	4	−2
1996	112	+8	64	12	−3	9	−24
1997	103	−1	1	19	+4	16	−4
1998	99	−5	25	26	+11	121	−55
Total	**832**	**0**	**120**	**120**	**0**	**184**	**−92**

$$\bar{x} = \frac{\Sigma x}{n} = \frac{832}{8} = 104; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{120}{8} = 15$$

Applying the formula, $$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{8 \times -92}{\sqrt{8 \times 120}\sqrt{8 \times 184}} = \frac{-92}{10.954 \times 13.564}$$

$$= \frac{-92}{148.580} = -0.619$$

Interpretation: Since the coefficient of correlation $r = -0.619$ is moderately negative, it indicates a moderately significant inverse correlation between the two variables. Hence, we conclude that as the production index increases, the number of unemployed decreases and vice versa.

Example 9.3: The following table gives the distribution of items of production and the relatively defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

Size-group	**15–16**	**16–17**	**17–18**	**18–19**	**19–20**	**20–21**
No. of items	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	114

Solution: Let group size be denoted by variable x and the number of defective items by variable y. Calculations for Karl Pearson's correlation coefficient are shown below:

Size Group	***Mid-value***	***dx=m−17.5***	d_x^2	***% of Defective Items (y)***	***dy=y−50***	d_y^2	d_x
15–16	15.5	−2	4	75	+25	625	−50
16–17	16.5	−1	1	60	+10	100	−10
17–18	17.5	0	0	50	0	0	0
18–19	18.5	+1	1	50	0	0	0
19–20	19.5	+2	4	45	−5	25	−10
20–21	20.5	+3	9	38	−12	144	−36
Total		**3**	**19**		**18**	**894**	**−106**

Substituting values in the formula of Karl Pearson's correlation coefficient r, we have

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{6\times -106 - 3\times 18}{\sqrt{6\times 19 - (3)^2}\sqrt{6\times 894 - (18)^2}} = \frac{-636-54}{\sqrt{105}\sqrt{5040}}$$

$$= -\frac{690}{727.46} = -0.949$$

Interpretation: Since the value of r is negative and is moderately close to -1, the statistical association between x (size group) and y (percent of defective items) is moderate and negative. We conclude that the number of faulty items decreases when the group size increases.

Example 9.4: The following data relate to employees' ages and the number of days they reported sick in a month.

Employees	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Age	30	32	35	40	48	50	52	55	57	61
Sick days	1	0	2	5	2	4	6	5	7	8

Calculate Karl Pearson's coefficient of correlation and interpret it.

Solution: To represent age and sick days by variables x and y, respectively. Calculations for the value of the correlation coefficient are shown below:

Age (x)	$d_x = x - \bar{x}$	d_x^2	***Sick days (y)***	$d_y = y - \bar{y}$	d_y^2	$d_x \cdot d_y$
30	-16	256	1	-3	9	48
32	-14	196	0	-4	16	56
35	-11	121	2	-2	4	22
40	-6	36	5	1	1	-6
48	2	4	2	-2	4	-4
50	4	16	4	0	0	0

52	6	36	6	2	4	12
55	9	81	5	1	1	9
57	11	121	7	3	9	33
61	15	225	8	4	16	60
Σ = 460	**0**	**1092**	**Σ = 40**	**0**	**64**	**230**

$$\bar{x} = \frac{\Sigma x}{n} = \frac{460}{10} = 46 \text{ and } \bar{y} = \frac{\Sigma y}{n} = \frac{40}{10} = 4$$

Substituting values in the formula of Karl Pearson's correlation coefficient *r*, we have

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10 \times 230}{\sqrt{10 \times 1092}\sqrt{10 \times 64}}$$

$$= \frac{230}{264.363} = 0.870$$

Interpretation: Since the value of r is positive, the age of employees and the number of sick days are highly correlated. Hence, we conclude that as an employee's age increases, he is likely to go on sick leave more often than others.

Example 9.5: A computer, while calculating the correlation coefficient between two variables *x* and *y* from 25 pairs of observations, obtained the following results:
$n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$ and $\Sigma y = 100$, $\Sigma y^2 = 460$, $\Sigma xy = 508$
It was, however, discovered at the time of checking that he had copied down two pairs of observations:

x	*y*
6	14
8	6

instead of

x	*y*
8	12
6	8

Obtain the correct value of the correlation coefficient between *x and y.*
Solution: The corrected values for the terms needed in the formula of Pearson's correlation coefficient are determined as follows:
Correct $\Sigma x = 125 - (6 + 8 - 8 - 6) = 125$
Correct $\Sigma y = 100 - (14 + 6 - 12 - 8) = 100$
Correct $\Sigma x^2 = 650 - \{(6)^2 + (8)^2 - (8)^2 - (6)^2\}$
$= 650 - \{36 + 64 - 64 - 36\} = 650$
Correct $\Sigma y^2 = 460 - \{(14)^2 + (6)^2 - (12)^2 - (8)^2\}$
$= 460 - \{196 + 36 - 144 - 64\} = 436$
Correct $\Sigma xy = 508 - \{(6 \times 14) + (8 \times 6) - (8 \times 12) - (6 \times 8)\}$
$= 508 - \{84 - 48 - 96 - 48\} = 520$
Applying the formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}}$$

$$= \frac{13,000 - 12,500}{\sqrt{625}\sqrt{900}} = \frac{500}{25 \times 30} = 0.667$$

Thus, the correct correlation coefficient value between *x* and *y* is 0.667.

Example 13.6: Calculate the coefficient of correlation from the following bivariate frequency distribution:

Sales Revenue (Rs in lakh)	*Advertising Expenditure (Rs in '000)*				*Total*
	5–10	*10–15*	*15–20*	*20–25*	
75–125	4	1	—	—	5
125–175	7	6	2	1	16
175–225	1	3	4	2	10
225–275	1	1	3	4	9
Total	13	11	9	7	40

Solution: Let advertising expenditure and sales revenue be represented by variables x and y, respectively. The calculations for the correlation coefficient are shown below:

Revenue y	*Mid-value* (m)	$x \rightarrow$ *Mid-value* (m) / d_x / d_y	*Advertising Expenditure* 5–10 (7.5, −1)	10–15 (12.5, 0)	15–20 (17.5, 1)	20–25 (22.5, 2)	*Total*, f	fd_y	fd_y^2	fd_xd_y
75–125	100	−2	(8) 4	(0) 1	(0) —	(0) —	5	−10	20	8
125–175	150	−1	(7) 7	(0) 6	(−2) 2	(−2) 1	16	−16	16	3
175–225	200	0	(0) 1	(0) 3	(0) 4	(0) 2	10	0	0	0
225–275	250	1	(−1) 1	(0) 1	(3) 3	(8) 4	9	9	9	10
		Total, f	13	11	9	7	$n = 40$	$\Sigma d_y = -17$	$\Sigma d_y^2 = 45$	$\Sigma fd_xd_y = 21$
		fd_x	−13	0	9	14	$\Sigma fd_x = 10$			
		fd_x^2	13	0	9	28	$\Sigma fd_x^2 = 50$			
		fd_xd_y	14	0	1	6	$\Sigma fd_xd_y = 21$			

where $d_x = (m - 12.5)/5$ and $d_y = (m - 200)/50$

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{n\Sigma fd_y^2 - (\Sigma fd_y)^2}} = \frac{40\times 21 - 10\times -17}{\sqrt{40\times 50 - (10)^2}\sqrt{40\times 45 - (-17)^2}}$$

$$= \frac{840 + 170}{\sqrt{1900}\sqrt{1511}} = \frac{1010}{1694.373} = 0.596$$

Interpretation: Since the value of r is positive, advertising expenditure and sales revenue are positively correlated to the extent of 0.596. Hence we conclude that as expenditure on advertising increases, the sales revenue also increases.

Example 9.6: The following table gives the frequency, according to the marks, obtained by 67 students in an intelligence test. Measure the degree of relationship between age and marks:

Test Marks	Age in years 18	19	20	21	Total
200–250	4	4	2	1	11
250–300	3	5	4	2	14
300–350	2	6	8	5	21
350–400	1	4	6	10	21
Total	10	19	20	18	67

Solution: Let the students' age and their marks be represented by variables x and y, respectively. Calculations for the correlation coefficient for this bivariate data are shown below:

y \ x	d_y \ d_x	18 / −1	19 / 0	20 / 1	21 / 2	Total, f	fd_y	fd_y^2	fd_xd_y
200–250	−1	(4) 4	(0) 4	(−2) 2	(−2) 1	11	−11	11	0
250–300	0	(0) 3	(0) 5	(0) 4	(0) 2	14	0	0	0
300–350	1	(−2) 2	(0) 6	(8) 8	(10) 5	21	21	21	16
350–400	2	(−2) 1	(0) 4	(12) 6	(40) 10	21	42	84	50
Total, f		10	19	20	18	n = 67	Σfd_y = 52	Σfd_y^2 = 116	Σfd_xd_y = 66
fd_x		−10	0	20	36	Σfd_x = 46			
fd_x^2		10	0	20	72	Σfd_x^2 = 102			
fd_xd_y		0	0	18	48	Σfd_xd_y = 66			

where $d_x = x - 19$, $d_y = (m - 275)/50$
Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{n\Sigma fd_y^2 - (\Sigma fd_y)^2}} = \frac{67\times 66 - 46\times 52}{\sqrt{67\times 102 - (46)^2}\sqrt{67\times 116 - (52)^2}}$$

$$= \frac{4422 - 2392}{\sqrt{6834 - 2116}\sqrt{7772 - 2704}} = \frac{2030}{\sqrt{4718}\sqrt{5068}}$$

$$= \frac{2030}{68.688\times 71.19} = 0.415$$

Interpretation: Since the value of r is positive, students' age and marks obtained in an intelligence test are positively correlated to the extent of 0.415. Hence, we conclude that students' intelligence test scores also increase as they age.

9.4 SUMMARY;

Correlation measures the strength and direction of a relationship between two variables. The coefficient of correlation quantifies this relationship on a scale from -1 to +1. Types of correlation include positive, negative, linear, and nonlinear. Analytical methods like the scatter diagram and Karl Pearson's coefficient help visually and numerically understand correlations.

Karl Pearson's method provides a precise value indicating the degree of linear association. The coefficient of determination explains how much variation in one variable is explained by the variation in another.

9.5 TECHNICAL TERMS:

- **Correlation:** A statistical measure that describes the strength and direction of a relationship between two variables.
- **Coefficient of Correlation:** A numerical value (ranging from -1 to +1) that quantifies the degree of linear relationship between two variables.
- **Positive Correlation**: A relationship where an increase in one variable tends to be associated with an increase in the other.
- **Negative Correlation:** A relationship where an increase in one variable tends to be associated with a decrease in the other.
- **Linear Correlation:** A correlation where the change in one variable is directly proportional to the change in another, forming a straight line in a graph.
- **Nonlinear Correlation:** A relationship where the variables are related but not in a straight-line manner.
- **Scatter Diagram**: A graphical representation of paired data points used to observe the nature of the relationship between two variables.
- **Karl Pearson's Correlation Coefficient:** A statistical formula used to compute the linear correlation coefficient (r) between two variables based on covariance and standard deviation.
- **Coefficient of Determination (R^2):** The square of the correlation coefficient; it indicates the proportion of variance in the dependent variable predictable from the independent variable.
- **Standard Deviation**: A measure of data points' dispersion or spread around a dataset's mean.

9.6 SELF-ASSESSMENT QUESTIONS:

1. Explain the meaning and significance of the term correlation.
2. What is meant by 'correlation'? Distinguish between positive, negative, and zero correlation.
3. What is correlation? Clearly explain its role with suitable illustrations from simple business problems
4. What is a scatter diagram? How does it help in studying the correlation between two variables, concerning both their direction and degree?
5. Find the correlation coefficient by Karl Pearson's method between x and y and interpret its value.

X	57	42	40	33	42	45	42	44	40	56	44	43
Y	10	60	30	41	29	27	27	19	18	19	31	29

6. Calculate the coefficient of correlation from the following data:

X	100	200	300	400	500	600	700
Y	30	50	60	80	100	110	130

7. The correlation coefficient between two variables, x and y, is 0.3. The covariance is 9. The variance of x is 16. Find the standard deviation of the y series.

8. Calculate Karl Pearson's correlation coefficient between age and playing habits from the data below. Comment on the value

Age	20	21	22	23	24	25
No. of students	500	400	300	240	200	160
Regular Players	400	300	180	96	60	24

9. Find the coefficient of correlation between age and the sum assured (in 1000 Rs) from the following table:

Age Group (Years)	***Sum Assured (in Rs)***				
	10	20	30	40	50
20-30	4	6	3	7	1
30-40	2	8	15	7	1
40-50	3	9	12	6	2
50-60	8	4	2	-	-

9.7 REFERENCES:

1. Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.
2. Levin, R.I., & Rubin, D.S. (2017). *Statistics for Management* (7th ed.). Pearson Education.
3. Sharma, J.K. (2020). *Business Statistics*. Pearson Education India.

Dr. G. Malathi

LESSON- 10

SPEARMAN'S RANK CORRELATION COEFFICIENT

OBJECTIVES:

The purpose of studying this lesson is:

- To understand the concept and importance of Spearman's Rank Correlation Coefficient.
- To identify situations where Spearman's Rank method is appropriate.
- To calculate the rank correlation when ranks are given.
- To assign ranks to raw data when ranks are not provided.
- To handle tied ranks and apply correction factors accurately.
- To understand the application of the least squares method to find the best fit line in a dataset.

STRUCTURE:

10.1 INTRODUCTION TO SPEARMAN'S RANK CORRELATION COEFFICIENT:

This method of finding the correlation coefficient between two variables was developed by the British psychologist Charles Edward Spearman in 1904. This method measures the association between two variables when only *ordinal* (*or rank*) *data* are available. In other words, this method is applied where a quantitative measure of certain qualitative factors, such as judgment, brand personalities, TV programs, leadership, color, and taste, cannot be fixed. Still, individual observations can be arranged in a definite order (called rank). The ranking is decided by using a set of ordinal rank numbers, with 1 for the individual observation ranked first in quantity or quality and n for the individual observation ranked last in a group of n pairs of observations. Mathematically, Spearman's rank correlation coefficient is defined as:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Where R = rank correlation coefficient

R_1 = rank of observations concerning the first variable

R_2 = rank of observations for the second variable
D = $R_1 - R_2$, difference in a pair of ranks
N = number of paired observations or individuals being ranked

The number '6' is placed in the formula as a scaling device. It ensures that R's possible range is from -1 to 1. While using this method, we may encounter three types of cases.

Advantages and Disadvantages of Spearman's Correlation Coefficient Method

Advantages

1. This method is easy to understand, and its application is more straightforward than Pearson's.
2. This method is helpful for correlation analysis when variables are expressed in qualitative terms like beauty, intelligence, honesty, efficiency, and so on.
3. This method is appropriate to measure the association between two variables if the data type is at least an ordinal scale (ranked)
4. The sample data of the values of two variables is converted into ranks, either in ascending order or descending order, to calculate the degree of correlation between them.

Disadvantages

1. The values of both variables are assumed to be normally distributed and describe a linear relationship rather than a nonlinear relationship.
2. Significant computational time is required when the number of pairs of values of two variables exceeds 30.
3. This method cannot be applied to measure the association between two variables with grouped data.

10.2 WHEN RANKS ARE GIVEN:

When observations in a data set are arranged in a particular order (rank), take the differences in pairs of observations to determine d. Square these differences and obtain the total Σd^2. Apply formula (13-4) to calculate the correlation coefficient.

Example 10.1: The rank correlation coefficient between debenture and share prices is 0.143. If the sum of the squares of the differences in ranks is 48, find the values of n.
Solution: The formula for Spearman's correlation coefficient is as follows:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Given, R = 0.143, $\Sigma d^2 = 48$ and n=7. Substituting values in the formula, we get

$$0.143 = 1 - \frac{6\times 48}{n(n^2-1)} = 1 - \frac{288}{n^3-n}$$

$$0.143\ (n^3 - n) = (n^3 - n) - 288$$

$$n^3 - n - 336 = 0 \quad \text{or} \quad (n-7)(n^2 + 7n + 48) = 0$$

This implies that either $n - 7 = 0$, that is, $n = 7$ or $n^2 + 7n + 48 = 0$. But $n^2 + 7n + 48 = 0$ on simplification gives the undesirable value of n because its discriminant $b^2 - 4ac$ is negative. Hence $n = 7$.

Example 10.2: The ranks of 15 students in two subjects, A and B, are below. The two numbers within brackets denote the ranks of a student in A and B subjects, respectively.
(1, 10), (2, 7), (3, 2), (4, 6), (5, 4), (6, 8), (7, 3), (8, 1), (9, 11), (10, 15), (11, 9), (12, 5), (13, 14), (14, 12), (15, 13)
Find Spearman's rank correlation coefficient.
Solution: Since the ranks of students for their performance in two subjects are given, calculations for the rank correlation coefficient are shown below:

Rank in A R1	***Rank in B R2***	***Difference d=R1−R2***	***d^2***
1	10	−9	81
2	7	−5	25
3	2	1	1
4	6	−2	4
5	4	1	1
6	8	−2	4
7	3	4	16
8	1	7	49
9	11	−2	4
10	15	−5	25
11	9	2	4
12	5	7	49
13	14	−1	1
14	12	2	4
15	13	2	4
			$\Sigma d^2 = 272$

Apply the formula, $R = 1 - \Sigma\frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 272}{15\{(15)^2 - 1\}}$

$$= 1 - \frac{1632}{3360} = 1 - 0.4857 = 0.5143$$

The result shows a moderate positive correlation between students' performance in the two subjects.

Example 10.3: An office has 12 clerks. The long-serving clerks feel they should have a seniority increment based on their length of service built into their salary structure. Their departmental manager and the personnel department assess their efficiency, producing a ranking of efficiency. This is shown below, together with a ranking of their length of service.

Ranking according to the length of service	1	2	3	4	5	6	7	8	9	10	11	12
Ranking according to efficiency	2	3	5	1	9	10	11	12	8	7	6	4

Do the data support the clerks' claim for a seniority increment?
Solution: Since ranks are already given, calculations for the rank correlation coefficient are shown below:

Rank According to Length of Service (R_1)	*Rank According to Efficiency (R_2)*	*Difference ($d = R_1 - R_2$)*	*d^2*
1	*2*	*-1*	*1*
2	*3*	*-1*	*1*
3	*5*	*-2*	*4*
4	*1*	*3*	*9*
5	*9*	*-4*	*16*
6	*10*	*-4*	*16*
7	*11*	*-4*	*16*
8	*12*	*-4*	*16*
9	*8*	*1*	*1*
10	*7*	*3*	*9*
11	*6*	*5*	*25*
12	*4*	*8*	*64*
			$\Sigma d^2 = 178$

Applying the formula, $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

$$= 1 - \frac{6 \times 178}{12(144-1)} = 1 - \frac{1068}{1716} = 0.378$$

The result shows a low-degree positive correlation between length of service and efficiency, so the clerks' claim for a seniority increment based on length of service is not justified.

Example 10.4: Ten competitors in a beauty contest are ranked by three judges in the following order:

Judge 1	1	6	5	10	3	2	4	9	7	8
Judge 2	3	5	8	4	7	10	2	1	6	9
Judge 3	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges have the nearest approach to familiar tastes in beauty.

Solution: The pair of judges who have the nearest approach to familiar taste in beauty can be obtained in $^3C_2 = 3$ ways as follows:

1. Judge 1 and Judge 2.
2. Judge 2 and Judge 3.
3. Judge 3 and Judge 1.

Calculations for comparing their ranking are shown below:

Judge 1 R_1	Judge 2 R_2	Judge 3 R_3	$d^2 = (R_1 - R_2)^2$	$d^2 = (R_2 - R_3)^2$	$d^2 = (R_3 - R_1)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
			$\Sigma d^2 = 200$	$\Sigma d^2 = 214$	$\Sigma d^2 = 60$

Applying the formula

$$R_{12} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 200}{10(100-1)} = 1 - \frac{1200}{990} = -0.212$$

$$R_{23} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 214}{10(100-1)} = 1 - \frac{1284}{990} = -0.297$$

$$R_{13} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 60}{10(100-1)} = 1 - \frac{360}{990} = 0.636$$

Since the correlation coefficient $R_{13} = 0.636$ is the largest, judges 1 and 3 have the closest approach to familiar tastes in beauty.

10.3 WHEN RANKS ARE NOT GIVEN

When pairs of observations in the data set are not ranked as in Case 1, the ranks are assigned by taking either the highest or lowest value as 1 for both variables.

Example 10.5: Quotations of index numbers of security prices of a particular joint stock company are given below:

Year	*Debenture Price*	*Share Price*
1	97.8	73.2
2	99.2	85.8
3	98.8	78.9
4	98.3	75.8
5	98.4	77.2
6	96.7	87.2
7	97.1	83.8

The rank correlation method determines the relationship between debenture and share prices.
Solution: Let us start ranking from the lowest value for both variables, as shown below:

Debenture Price (x)	Rank	Share Price (y)	Rank	Difference $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
97.8	3	73.2	1	2	4
99.2	7	85.8	6	1	1
98.8	6	78.9	4	2	4
98.3	4	75.8	2	2	4
98.4	5	77.2	3	2	4
96.7	1	87.2	7	– 6	36
97.1	2	83.8	5	– 3	9
					$\Sigma d^2 = 62$

Applying the formula $R = 1 - \frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 62}{(7)^3 - 7}$

$$= 1 - \frac{372}{336} = 1 - 0.107 = -0.107$$

The result shows a low negative correlation between a specific joint stock company's debenture and share prices.

Example 10.6: An economist wanted to find out if there was any relationship between the unemployment rate in a country and its inflation rate. Data gathered from 7 countries for the year 2004 is given below:

Country	*Unemployment Rate (Percent)*	*Inflation Rate (Percent)*
A	4.0	3.2
B	8.5	8.2
C	5.5	9.4
D	0.8	5.1
E	7.3	10.1
F	5.8	7.8
G	2.1	4.7

Find the degree of linear association between a country's unemployment rate and its level of inflation.

Solution: Let us start ranking from the lowest value for both variables, as shown below:

Unemployment Rate (x)	Rank R_1	Inflation Rate (y)	Rank R_2	Difference $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
4.0	3	3.2	1	2	4
8.5	7	8.2	5	2	4
5.5	4	9.4	6	– 2	4
0.8	1	5.1	3	– 2	4
7.3	6	10.1	7	– 1	1
5.8	5	7.8	4	1	1
2.1	2	4.7	2	0	0
					$\Sigma d^2 = 18$

Applying the formula,

$$R = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \times 18}{(7)^3 - (7)} = 1 - \frac{108}{336} = 0.678$$

The result shows a moderately high positive correlation between the seven countries' unemployment and inflation rates.

10.4 WHEN RANKS ARE REPEATING:

While ranking observations in the data set by taking either the highest value or the lowest value as rank 1, we may encounter a situation of more than one observation being of equal size. In such a case, the rank assigned to individual observations is an average of the ranks these observations would have gotten had they differed. For example, if two observations are ranked equal in third place, then the average rank of (3 + 4)/2 = 3.5 is assigned to these two observations. Similarly, if three observations are ranked equal in third place, then the average rank of (3 + 4 + 5)/3 = 4 is assigned to these three observations.

While equal ranks are assigned to a few observations in the data set, an adjustment is made in the Spearman rank correlation coefficient formula as given below:

$$R = 1 - \frac{6\left\{\sum d^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots\right\}}{n(n^2 - 1)}$$

Where m_i (i = 1, 2, 3,…) represents the number of times an observation is repeated in the data set for both variables.

Example 10.7: A financial analyst wanted to determine whether inventory turnover influences any company's earnings per share (in percent). A random sample of 7 companies listed on a stock exchange was selected, and the following data were recorded for each.

Company	*Inventory Turnover (Number of Times)*	*Earnings per Share (Percent)*
A	4	11
B	5	9
C	7	13
D	8	7
E	6	13
F	3	8
G	5	8

Find the strength of the association between inventory turnover and earnings per share. Interpret this finding.

Solution: Let us start ranking from the lowest value for both variables. Since there are tied ranks, the sum of the tied ranks is averaged and assigned to each tied observation, as shown below.

Inventory Turnover (x)	*Rank* R_1	*Earnings Per Share (y)*	*Rank* R_2	*Difference* $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
4	2	11	5	– 3.0	9.00
5	3.5	9	4	– 0.5	0.25
7	6	13	6.5	0.5	0.25
8	7	7	1	6.0	36.00
6	5	13	6.5	– 1.5	2.25
3	1	8	2.5	– 1.5	2.25
5	3.5	8	2.5	1.0	1.00
					$\Sigma d^2 = 51$

It may be noted that a value 5 of variable x is repeated twice (m1 = 2), and values 8 and 13 of variable y are also repeated twice, so $m_2 = 2$ and $m_3 = 2$. Applying the formula:

$$R = 1 - \frac{6\left\{\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{51 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right\}}{7(49-1)}$$

$$= 1 - \frac{6\{51 + 0.5 + 0.5 + 0.5\}}{336} = 1 - 0.9375 = 0.0625$$

The result shows a weak positive association between inventory turnover and earnings per share.

Example 10.8: Obtain the rank correlation coefficient between the variables *x* and *y* from the following observed values.

X	50	55	65	50	55	60	50	65	70	75
Y	110	110	115	125	140	115	130	120	115	160

Solution: Let us start ranking from the lowest value for both variables. Moreover, certain observations in both sets of data are repeated, so the ranking is done by a suitable average value, as shown below.

Variable x	*Rank* R_1	*Variable* y	*Rank* R_2	*Difference* $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
50	2	110	1.5	0.5	0.25
55	4.5	110	1.5	3.0	9.00
65	7.5	115	4	3.5	12.25
50	2	125	7	– 5.0	25.00
55	4.5	140	9	– 4.5	20.25
60	6	115	4	2.0	4.00
50	2	130	8	– 6.0	36.00
65	7.5	120	6	1.5	2.25
70	9	115	4	5.0	25.00
75	10	160	10	0.0	00.00
					$\Sigma d^2 = 134.00$

It may be noted that for variable *x*, 50 is repeated thrice ($m_1 = 3$), 55 is repeated twice ($m_2 = 2$), and 65 is repeated twice ($m_3 = 2$). Also, for variable *y*, 110 is repeated twice ($m_4 = 2$) and 115 thrice ($m_5 = 3$). Applying the formula:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \frac{1}{12}(m_4^3 - m_4) + \frac{1}{12}(m_5^3 - m_5)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{134 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{10(100-1)}$$

$$= 1 - \frac{6\,[134 + 2 + 0.5 + 0.5 + 0.5 + 2]}{990} = 1 - \frac{6 \times 139.5}{990} = 1 - \frac{837}{990}$$

$$= 1 - 0.845 = 0.155$$

The result shows a weak positive association between variables *x* and *y*

10.5 METHOD OF LEAST SQUARES:

The least squares method to calculate the correlation coefficient requires the values of regression coefficients *by and by*.

$$r = \sqrt{b_{xy} \times b_{yx}}$$

In other words, the correlation coefficient is the geometric mean of two regression coefficients.

10.6 SUMMARY:

Spearman's Rank Correlation Coefficient is a non-parametric measure used to determine the strength and direction of association between two ranked variables. It evaluates how well the relationship between two data sets can be described using a monotonic function, making it suitable for ordinal data or data where numerical ranks are assigned. When the ranks are provided, the formula can be directly applied by calculating the differences between the paired ranks. When ranks are not given, the raw data must first be ranked for each variable before proceeding with the calculation. If some values are repeated in the data (ties), they are assigned the average of the ranks they would have otherwise occupied. A correction factor is included in the formula to address the impact of tied ranks. The Method of Least Squares is a statistical approach used to determine the line or curve that best fits a set of data points by minimizing the total of the squared differences between the observed and predicted values.

10.7 TECHNICAL TERMS:

- **Rank:** The position of a data value in an ordered list used to replace raw scores in non-parametric methods.
- **Tied Ranks:** Occurs when two or more data values are the same; they are assigned the average of the ranks they occupy.
- **Monotonic Relationship:** A relationship where the variables move consistently in the same or opposite direction, though not necessarily at a constant rate.
- **Correction Factor:** An adjustment added to Spearman's formula when tied ranks are present, ensuring accurate coefficient computation.

- **Least Squares Method:** A technique to determine the line of best fit by minimizing the sum of the squares of the deviations (errors) from the observed values to the predicted values.

10.8 SELF-ASSESSMENT PROBLEMS:

1. What is the coefficient of rank correlation? Bring out its usefulness. How does this coefficient differ from the coefficient of correlation?
2. What is Spearman's rank correlation coefficient? How does it differ from Karl Pearson's coefficient of correlation?
3. The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was 0.2. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct coefficient of rank correlation.
4. The ranking of 10 students following their performance in two subjects, A and B, are as follows:

A	6	5	3	10	2	4	9	7	8	1
B	3	8	4	9	1	6	10	7	5	2

Calculate the rank correlation coefficient and comment on its value.

5. A firm examined eight applicants for a clerical post. Calculate the rank correlation coefficient from the marks obtained by the applicants in the accountancy and statistics papers.

Applicant	A	B	C	D	E	F	G	H
Marks in Accountancy	15	20	28	12	40	60	20	80
Marks in Statistics	40	30	50	30	20	10	30	60

6. Seven methods of imparting business education were ranked by the MBA students of two universities as follows:

Method of Teaching	1	2	3	4	5	6	7
Ranking by students of University A	2	1	5	3	4	7	6
Ranking by students of University B	1	3	2	4	7	5	6

Calculate the rank correlation coefficient and comment on its value.

10.9 REFERENCES:

1. Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.
2. Levin, R.I., & Rubin, D.S. (2017). *Statistics for Management* (7th ed.). Pearson Education.
3. Sharma, J.K. (2020). *Business Statistics*. Pearson Education India.

Dr. G. Malathi

LESSON-11 & 12

REGRESSION

OBJECTIVES:

The purpose of studying this chapter is:

- To understand the concept and significance of the coefficient of correlation.
- To distinguish between different types of correlations (positive, negative, linear, and non-linear).
- To construct and interpret scatter diagrams to observe data relationships.
- To calculate Karl Pearson's correlation coefficient and interpret its value.
- To understand the concept of the coefficient of determination and its role in explaining variability.

STRUCTURE:

11.1 INTRODUCTION TO REGRESSION:

In Lesson 9, we introduced the concept of statistical relationship between two variables, such as the level of sales and the amount of advertising, the yield of a crop and the amount of fertilizer used, the price of a product and its supply, and so on. The relationship between such variables indicates the degree and direction of their association but fails to answer the following question:

- Is there any functional (or algebraic) relationship between two variables? If yes, can it be used to estimate the most likely value of one variable, given the value of other variables?

The statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable based on the given value of another variable is called *regression analysis*. The variable whose value is estimated using the algebraic equation is called *the dependent (or response) variable, and the variable whose value is used to estimate this value is called the independent* (*regressor* or *predictor*) *variable.* The linear algebraic equation for expressing a dependent variable in terms of the independent variable is called a *linear regression equation*.

The term regression was used in 1877 by Sir Francis Galton while studying the relationship between the height of fathers and their sons. He found that though the 'tall father has tall sons,' the average height of the tall father is *x* above the general height, and the average height of sons is 2*x*/3 above the general height. Galton described such a fall in the average height as a regression to mediocrity. However, Galton's theory is not universally applicable, and the term regression is applied to other variables in business and economics. The term regression in the literary sense is also called 'moving backward.'

The basic differences between correlation and regression analysis are summarized as follows:

1. Developing an algebraic equation between two variables from sample data and predicting the value of one variable, given the value of the other variable, is referred to as regression analysis, while measuring the strength (or degree) of the relationship between two variables is referred to as correlation analysis. The sign of the correlation coefficient indicates the nature (direct or inverse) of the relationship between two variables. In contrast, the correlation coefficient's absolute value indicates the extent of the relationship.
2. Correlation analysis determines an association between two variables*, x, and y,* but not that they have a cause-and-effect relationship. In contrast to correlation, regression analysis determines the cause-and-effect relationship between *x* and *y*; that is, a change in the independent variable x causes a corresponding change (effect) in the value of the dependent variable *y* if all other factors that affect *y* remain unchanged.
3. In linear regression analysis, one variable is considered the dependent variable and the other the independent variable, while in correlation analysis, both variables are considered independent.
4. *The coefficient of determination r^2 indicates the proportion of total variance in the dependent variable explained or accounted for by the variation in the independent variable*. Since the value of r^2 is determined from a sample, it is subject to sampling error. Even if the value of r^2 is high, the assumption of a linear regression may be incorrect because it may represent a portion of the relationship in the form of a curve.

11.1.1 Advantages of Regression Analysis

The following are some crucial advantages of regression analysis:

1. Regression analysis helps develop a regression equation by which the value of a dependent variable can be estimated given the value of an independent variable.

2. Regression analysis helps determine the standard error of estimate to measure the variability or spread of values of a dependent variable concerning the regression line. The smaller the variance and error of estimate, the closer the pair of values (x, y) falls about the regression line and the better the line fits the data. That is, a reasonable estimate can be made of variable y's value. When all the points fall on the line, the standard error of estimate equals zero.

3. When the sample size is large ($df \geq 29$), the interval estimation for predicting the value of a dependent variable based on the standard error of the estimate is considered acceptable by changing the values of either x or y. The magnitude of r^2 remains the same regardless of the values of the two variables.

11.2 TYPES OF REGRESSION MODELS:

The primary objective of regression analysis is to develop a regression model to explain the association between two or more variables in the population. A regression model is a mathematical equation that predicts the value of the dependent variable based on the known values of one or more independent variables.

The particular form of the regression model depends on the nature of the problem under study and the available data type. However, an equation relating a dependent variable to one or more independent variables can describe every association or relationship.

Simple and Multiple Regression Models
If a regression model characterizes the relationship between a dependent variable, y, and only one independent variable, x, then such a regression model is called a *simple regression model*.

However, if more than one independent variable is associated with a dependent variable, such a regression model is called a *multiple regression model*. For example, sales turnover of a product (a dependent variable) is associated with various independent variables such as product price, expenditure on advertisement, product quality, competitors, and so on. Now, if we want to estimate possible sales turnover concerning only one of these independent variables, it is an example of a simple regression model; otherwise, a multiple regression model is applicable.

Linear and Nonlinear Regression Models
If the value of a dependent (response) variable y in a regression model tends to increase in direct proportion to an increase in the values of the independent (predictor) variable x, then such a regression model is called a *linear model*. Thus, it can be assumed that the mean value of the variable y for a given value of x is related by a straight-line relationship. Such a *relationship is called a simple linear regression model* expressed in terms of the population parameters β_0 and β_1 as:

$$E(y|x) = \beta_0 + \beta_1 x$$

where β_0 = y-intercept that represents the mean (or average) value of the dependent variable y when $x = 0$

β_1 =slope of the regression line that represents the expected change in the value of y (either positive or negative) for a unit change in the value of x.

11.3 ESTIMATION: THE METHOD OF LEAST SQUARES:

To estimate the values of regression coefficients β_0 and β_1, suppose a sample of n pairs of observations (x_1, y_1), (x_2, y_2),…, (x_n, y_n) is drawn from the population under study. A method that provides the best linear unbiased estimates of β_0 and β_1 is called the *method of least squares*. The forecast of β_0 and β_1 should result in a straight line that is the *'best fit'* to the data points. The straight line so drawn is referred to as *the 'best fitted' (least squares or estimated) regression line because the sum of the squares of the vertical deviations (difference between*

the actual values of y and the estimated values predicted from the fitted line) is as small as possible.

We may express the given n observations in the sample data as:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{or} \quad e_i = y_i - (\beta_0 + \beta_1 x_i), \text{ for all } i$$

Mathematically, we intend to minimize

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Let b_0 and b_1 be the least-squares estimators of β_0 and β_1, respectively. The least-squares estimators b_0 and b_1 must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{b_0 b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{b_0 b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0$$

After simplifying these two equations, we get

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

The above equations are called the *least-squares normal equations*. Solving those two equations yields the least-squares estimators b0 and b1 values.
Hence, the *fitted* or *estimated regression line* is given by:

$$\hat{y} = b_0 + b_1 x$$

The fitted value is also called the *predicted value of y* because if the actual value of y is unknown, then it would be expected for a given value of x using the estimated regression line.
Remark: The sum of the residuals is zero for any least-squares regression line. Since $\Sigma y_i = \Sigma \hat{y}_i$, therefore so $\Sigma e_i = 0$.

11.3.1 Assumptions for a Simple Linear Regression Model

To make a valid statistical inference using regression analysis, we make certain assumptions about the bivariate population from which a sample of paired observations is drawn and how observations are generated. These assumptions form the basis for the application of simple linear regression models.

Assumptions

1. The relationship between the dependent variable y and the independent variable x exists and is linear. The average relationship between x and y can be described by a simple linear regression equation $y = a + bx + e$, where e is the deviation of a particular y value from its expected value for a given value of the independent variable x.
2. For every value of the independent variable x, there is an expected (or mean) value of the dependent variable y, which is usually distributed. The mean of these normally distributed values falls on the line of regression.

3. The dependent variable *y* is a continuous random variable, whereas the independent variable x's values are fixed and not random.
4. The sampling error associated with the expected value of the dependent variable *y* is assumed to be an independent random variable distributed generally with mean zero and constant standard deviation. The errors are not related to each other in successive observations.
5. The standard deviation and variance of expected values of the dependent variable *y* about the regression line are constant for all values of the independent variable *x* within the range of the sample data.
6. The value of the dependent variable cannot be estimated for the value of an independent variable lying outside the range of values in the sample data.

11.3.2 Parameters of a Simple Linear Regression Model

The fundamental aim of regression analysis is to determine a regression equation (line) that makes sense and fits the representative data such that the variance error is as small as possible. This implies that the regression equation should be used adequately for prediction. J. R. Stockton stated that

- *The device used for estimating the values of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables. Such a line is called the line of regression*.

The two variables, *x* and *y*, which are correlated, can be expressed in terms of each other in the form of straight-line equations called *regression equations*. Such lines should provide the best fit for sample data and population data. The algebraic expression of regression lines is written as:

- The regression equation of *y* on *x*

$$y = a + bx$$

is used to estimate the value *of y* for given values of *x*.

- Regression equation of *x* on *y*

$$x = c + dy$$

is used to estimate the value of *x* for given values of *y*.

Remarks

1. When variables *x* and *y* correlate perfectly (positive or negative), these lines coincide; we have only one line.
2. The higher the degree of correlation, the nearer the two regression lines are to each other.
3. The lower the degree of correlation, the farther apart the two regression lines are. That is, when $r = 0$, the two lines are at right angles.
4. Two linear regression lines intersect at the point of the average value of variables *x* and *y*.

11.3.3 Regression Coefficients

To estimate the values of population parameters β_0 and β_1, under certain assumptions, the fitted or estimated regression equation representing the straight-line regression model is written as:

$$\hat{y} = a + bx$$

where $\hat{y}$ = estimated average (mean) value of dependent variable *y* for a given value of independent variable *x*.

a or b_0 = *y*-intercept that represents the average value of $\hat{y}$

B = slope of the regression line that represents the expected change in the value of *y* for a unit change in the value of *x*

To determine the value of a given value of x, this equation requires the determination of two unknown constants, *a (intercept) and b (also called the* regression coefficient). Once these constants are calculated, the regression line can compute an estimated value of the dependent variable y for a given value of the independent variable x.
The particular values of a and b define a specific linear relationship between x and y based on sample data. The coefficient 'a' represents the *level of the fitted line* (i.e., the distance of the line above or below the origin) when x equals zero, whereas coefficient 'b' represents the *slope of the line* (a measure of the change in the estimated value of y for a one-unit change in x).
The regression coefficient *'b'* is also denoted as:

- b_{yx} (*regression coefficient of y on x*) in the regression line, $y = a + bx$
- b_{xy} (*regression coefficient of x on y*) in the regression line, $x = c + dy$

11.3.4 Properties of Regression Coefficients

1. The correlation coefficient is the geometric mean of two regression coefficients, that is, $r = \sqrt{b_{yx} \times b_{xy}}$.
2. If one regression coefficient is more significant than the other, then the other regression coefficient must be less than one because the value of the correlation coefficient r cannot exceed one. However, both the regression coefficients may be less than one.
3. Both regression coefficients must have the same sign (either positive or negative). This property rules out the opposite signs of the two regression coefficients.
4. The correlation coefficient will have the same sign (either positive or negative) as that of the two regression coefficients. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then $r = -\sqrt{0.664 \times 0.234} = -0.394$.
5. The arithmetic mean of regression coefficients b_{xy} and b_{yx} is more than or equal to the correlation coefficient r, that is, $(b_{yx} + b_{xy})/2 \geq r$. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, the arithmetic mean of these two values is $(-0.664 - 0.234)/2 = -0.449$, more than the $r = -0.394$ value.
6. Regression coefficients are independent of origin but not of scale.

11.3.5 Methods to Determine Regression Coefficients
The following are the methods to determine the parameters of a fitted regression equation.

Least Squares Normal Equations
Let $\hat{y} = a + bx$ be the least squares line of y on x, where $\hat{y}$ -is the estimated average value of the dependent variable y. The line that minimizes the sum of squares of the deviations of the observed values of y from those predicted is the best-fitting line. Thus, the sum of residuals for any least-squares line is minimum, were

$$L = \Sigma (y - \hat{y})^2 = \Sigma \{y - (a + bx)\}^2; \quad a, b = \text{constants}$$

Differentiating L concerning a and b and equating to zero, we have

$$\frac{\partial L}{\partial a} = -2\Sigma\{y - (a + bx)\} = 0$$

$$\frac{\partial S}{\partial b} = -2\Sigma\{y - (a + bx)\}x = 0$$

Solving these two equations, we get the same set of equations as

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Where *n* is the number of pairs of x and y values in a sample of data, the above equations are called *normal equations* for the regression line of *y* on *x*. After solving these equations for *a* and *b*, the values of *a* and *b* are substituted in the regression equation, $y = a + bx$.
Similarly, if we have a least squares line $\hat{x} = c + dy$ of *x* on *y*, where $\hat{x}$ is the estimated mean value of the dependent variable *x*, then the normal equations will be

$$\Sigma x = nc + d\Sigma y$$
$$\Sigma xy = n\Sigma y + d\Sigma y^2$$

These equations are solved in the same manner as described above for constants *c* and *d*.
The values of these constants are substituted into the regression equation $x = c + dy$.

An alternative method to calculate the value of constants

Instead of using the algebraic method to calculate the values of *a* and *b*, we may directly use the results of the solutions of this normal equation.
The gradient *'b'* (regression coefficient of *y* on *x*) and *'d'* (regression coefficient of *x* on *y*) are calculated as:

$$b = \frac{S_{xy}}{S_{xx}}, \quad \text{where} \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$\text{and} \quad d = \frac{S_{yx}}{S_{yy}}, \quad \text{where} \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y\right)^2$$

Since the regression line passes through the point $(\bar{x}, \bar{y})$, the mean values of *x* and *y* and the regression equations can be used to find the value of constants *a* and *c* as follows:

$$a = \bar{y} - b\bar{x} \text{ for regression equation of } y \text{ on } x$$
$$c = \bar{x} - d\bar{y} \text{ for regression equation of } x \text{ on } y$$

The calculated values of *a*, *b*, and c, d are substituted in the regression line y = a + bx and x = c + dy, respectively, to determine the exact relationship.

Example 11.1: Use the least squares regression line to estimate the increase in sales revenue expected from an increase of 7.5 percent in advertising expenditure.

Firm	Annual % Increase in Advertising Expenditure	Annual % Increase in Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

Solution: Assume sales revenue (y) depends on advertising expenditure (x). Calculations for the regression line using the following normal equations are shown in Table 11.1

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Table 11.1: Calculation for Normal Equations

Sales Revenue (y)	Advertising Expenditure(x)	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
Σy=40	Σx=56	Σx^2=524	Σxy=373

Approach 1 (*Normal Equations*):

$$\Sigma y = na + b\Sigma x \quad \text{or} \quad 40 = 8a + 56b$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{or} \quad 373 = 56a + 524b$$

Solving these equations, we get

$a = 0.072$ and $b = 0.704$

Substituting these values in the regression equation

$y = a + bx = 0.072 + 0.704x$

For x = 7.5% or 0.075 increase in advertising expenditure, the estimated increase in sales revenue will be

$y = 0.072 + 0.704\ (0.075) = 0.1248$ or 12.48%

***Approach 2* (*Short-cut method*):**

$$b = \frac{S_{xy}}{S_{xx}} = \frac{93}{132} = 0.704,$$

$$\text{where } S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 373 - \frac{40 \times 56}{8} = 93$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 524 - \frac{(56)^2}{8} = 132$$

The intercept 'a' on the y-axis is calculated as:

$$a = \bar{y} - b\bar{x} = \frac{40}{8} - 0.704 \times \frac{56}{8} = 5 - 0.704 \times 7 = 0.072$$

Substituting the values of $a = 0.072$ and $b = 0.704$ in the regression equation, we get

$y = a + bx = 0.072 + 0.704\ x$

For $x = 0.075$, we have $y = 0.072 + 0.704\ (0.075) = 0.1248$ or 12.48%.

Example 11.2: The owner of a small garment shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in Rs 1000s and analyzed the results.

Week	1	2	3	4	5	6
Sales	2.69	2.62	2.80	2.70	2.75	2.81

Fit a linear regression equation to suggest to him the weekly rate at which his sales are rising and use this equation to estimate expected sales for the seventh week.

Solution: Assume sales (y) depend on weeks (x). Then the standard equations for the regression equation, $y = a + bx$, are written as:

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Calculations for sales during various weeks are shown in Table 11.2

Table 11.2: Calculations of Normal Equations

Week (x)	***Sales (y)***	x^2	***Xy***
1	2.69	1	2.69
2	2.62	4	5.24
3	2.80	9	8.40
4	2.70	16	10.80
5	2.75	25	13.75
6	2.81	36	16.86
Σx=21	**Σy=16.37**	**Σx^2=91**	**Σxy=57.74**

The gradient 'b' is calculated as:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{0.445}{17.5} = 0.025; \quad S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 57.74 - \frac{21 \times 16.37}{6} = 0.445$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 91 - \frac{(21)^2}{6} = 17.5$$

The intercept 'a' on the y-axis is calculated as

$$a = \bar{y} - b\bar{x} = \frac{16.37}{6} - 0.025 \times \frac{21}{6}$$

$$= 2.728 - 0.025 \times 3.5 = 2.64$$

Substituting the values $a = 2.64$ and $b = 0.025$ in the regression equation, we have

$$y = a + bx = 2.64 + 0.025x$$

For $x = 7$, we have $y = 2.64 + 0.025\ (7) = 2.815$

Hence, the expected sales during the seventh week will likely be Rs 2.815 (in Rs 1000s)

12.1 DEVIATIONS METHOD:

Calculations to the least squares normal equations become lengthy and tedious when the values of x and y are large. Thus, the following two methods may reduce the computational time.

(a) **Deviations Taken from Actual Mean Values of *x* and *y*.** If deviations of actual values of variables x and y are taken from their mean values $\bar{x}\,\bar{y}$, then the regression equations can be written as:

- Regression equation of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where b_{yx} = the regression coefficient of y on x
The value of b_{yx} can be calculated using the formula

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where b_{xy} = regression coefficient of x on y.
The value of b_{xy} can be calculated using a formula.

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}$$

(b) **Deviations Taken from Assumed Mean Values for *x* and *y*.** If the mean value of either x or y or both is in fractions, then we must prefer to take the deviations of the actual values of variables x and y from their assumed means.

- Regression equation of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{where } b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2}$$

n = number of observations
$d_x = x - A$; A is assumed mean of x
$d_y = y - B$; B is assumed to be the mean of y

- Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\text{where } b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_y^2 - (\Sigma d_y)^2}$$

n = number of observations
$d_x = x - A$; A is assumed mean of x
$d_y = y - B$; B is assumed to be the mean of y

(c) **Regression Coefficients in Terms of Correlation Coefficient:** If deviations are taken from actual mean values, then the values of regression coefficients can be alternatively calculated as follows:

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\text{Covariance}(x,y)}{\sigma_x^2} = r\cdot\frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{\text{Covariance}(x,y)}{\sigma_y^2} = r\cdot\frac{\sigma_x}{\sigma_y}$$

Example 12.1: The following data relate to the scores obtained by nine salesmen of a company in an intelligence test and their weekly sales (in Rs 1000)

Salesmen	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Test Scores	50	60	50	60	80	50	80	40	70
Weekly Sales	30	60	40	50	60	30	70	50	60

a) Obtain the regression equation of sales on intelligence test scores of the salespeople.
b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales?

Solution: Assume weekly sales (y) as the dependent variable and test scores (x) as the independent variable. Calculations for the following regression equation are shown in Table 11.3

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Table 12.1: Calculation for Regression Equation

Weekly Sales, x	***dx = x – 60***	***dx²***	***Test Score, y***	***dy = y – 50***	***dy²***	***dx·dy***
50	-10	100	30	-20	400	200
60	0	0	60	10	100	0
50	-10	100	40	-10	100	100
60	0	0	50	0	0	0
80	20	400	60	10	100	200
50	-10	100	30	-20	400	200
80	20	400	70	20	400	400
40	-20	400	50	0	0	0
70	10	100	60	10	100	100
$\Sigma x = 540$	**0**	**$\Sigma dx^2 = 1600$**	**$\Sigma y = 450$**	**0**	**$\Sigma dy^2 = 1600$**	**$\Sigma dx \cdot dy = 1200$**

$$\text{(a)}\quad \bar{x} = \frac{\Sigma x}{n} = \frac{540}{9} = 60; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{450}{9} = 50$$

$$b_{yx} = \frac{\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{1200}{1600} = 0.75$$

Substituting values in the regression equation, we have

$$y - 50 = 0.75(x - 60) \text{ or } y = 5 + 0.75x$$

For the test score $x = 65$ of the salesmen, we have

$$y = 5 + 0.75(65) = 53.75$$

Hence, we conclude that the weekly sales are expected to be Rs 53.75 (in Rs 1000s) for a test score of 65.

Example 12.2: A company is introducing a job evaluation scheme in which points for skill, responsibility, and so on are assigned to all jobs. Monthly pay scales (Rs in 1000s) are drawn up according to the number of points allocated and other factors such as experience and local conditions. To date, the company has applied this scheme to 9 jobs:

Job	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Points	5	25	7	19	10	12	15	28	16
Pay (Rs.)	3.0	5.0	3.25	6.5	5.5	5.6	6.0	7.2	6.1

Find the least squares regression line linking pay scales to points.

1. Estimate the monthly pay for a job graded by 20 points.

Solution: Assume monthly pay (y) as the dependent variable and job grade points (x)as the independent variable. Calculations for the following regression equation are shown in the table below.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Table 12.2: Calculation for Regression Equation

Grade Points, x	***dx = x – 15***	***dx²***	***Pay Scale, y***	***dy = y – 5***	***dy²***	***dx·dy***
5	-10	100	3.0	-2.0	4	20
25	10	100	5.0	0	0	0
7	-8	64	3.25	-1.75	3.06	14
19	4	16	6.5	1.50	2.25	6
10	-5	25	5.5	0.50	0.25	-2.5
12	-3	9	5.6	0.60	0.36	-1.8
15	0	0	6.0	1.00	1.00	0
28	13	169	7.2	2.2	4.84	28.6
16	1	1	6.1	1.1	1.21	1.1
Σx = 137	**Σdx = 2**	**Σdx²=484**	**Σy = 48.15**	**Σdy=3.15**	**Σdy²=16.97**	**Σdx·dy=65.40**

$$\bar{x} = \frac{\Sigma x}{n} = \frac{137}{9} = 15.22; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{48.15}{9} = 5.35$$

Since the mean values are non-integer, deviations are taken from the assumed mean, as shown below.

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9 \times 65.40 - 2 \times 3.15}{9 \times 484 - (2)^2} = \frac{582.3}{4352} = 0.133$$

Substituting values in the regression equation, we have

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad \text{or} \quad y - 5.35 = 0.133\,(x - 15.22) = 3.326 + 0.133x$$

(b) For job grade point $x = 20$, the estimated average pay scale is given by

$$y = 3.326 + 0.133x = 3.326 + 0.133\,(20) = 5.986$$

Hence, the likely monthly pay for a job with a grade point of 20 is Rs 5986.

Example 12.3: The following data give the ages and blood pressure of 10 women.

Age	56	42	36	47	49	42	60	72	63	55
Blood Pressure	147	125	118	128	145	140	155	160	149	150

1. Find the correlation coefficient between age and blood pressure.
2. Determine the least squares regression equation of blood pressure on age.
3. Estimate the blood pressure of a woman whose age is 45 years.

Solution: Assume blood pressure *(y)* is the dependent variable and age *(x)* is the independent variable. Calculations for the regression equation of blood pressure on age are shown in the following table.

Table 12.3: Calculation for Regression Equation

Age, x	*dx = x – 49*	*dx²*	*Blood, y*	*dy = y – 145*	*dy²*	*dx·dy*
56	7	49	147	2	4	14
42	-7	49	125	-20	400	140
36	-13	169	118	-27	729	351
47	-2	4	128	-17	289	34
49	0	0	145	0	0	0
42	-7	49	140	-5	25	35
60	11	121	155	10	100	110
72	23	529	160	15	225	345
63	14	196	149	4	16	56
55	6	36	150	5	25	30
Σx = 522	**Σdx = 32**	***Σdx²=1202***	**Σy = 1417**	***Σdy=-33***	***Σdy²=1813***	***Σdx·dy=1115***

(a) The coefficient of correlation between age and blood pressure is given by

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{10(1115) - (32)(-33)}{\sqrt{10(1202) - (32)^2}\sqrt{10(1813) - (-33)^2}}$$

$$= \frac{11150 + 1056}{\sqrt{12020 - 1024}\sqrt{18130 - 1089}} = \frac{12206}{13689} = 0.892$$

There is a high degree of positive correlation between age and blood pressure.

(b) The regression equation of blood pressure on age is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{522}{10} = 52.2; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{1417}{10} = 141.7$$

and

$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10(1115) - 32(-33)}{10(1202) - (32)^2} = \frac{12206}{10996} = 1.11$$

Substituting these values in the above equation, we have

$y - 141.7 = 1.11\ (x - 52.2)$ or $y = 83.758 + 1.11x$

This is the required regression equation of y on x.

(c) For women whose age is 45, the estimated average blood pressure will be

$y = 83.758 + 1.11(45) = 83.758 + 49.95 = 133.708$

Hence, the likely blood pressure of a woman of 45 years is 134.

Example 12.4: The General Sales Manager of Kiran Enterprises—an enterprise specializing in selling ready-made men's wear—is considering increasing its sales target to Rs 80,000. Upon reviewing the sales records for the past 10 years, it was discovered that the annual sales revenue and advertising expenditures were highly correlated, with a coefficient of 0.8. It was also observed that the average annual sales amounted to Rs 45,000, while the average annual advertising expenditure was Rs 30,000, with variances of Rs 1600 and Rs 625 in advertising expenditure, respectively.

Given the above, how much advertising expenditure would you suggest the General Sales Manager of the enterprise incur to meet his sales target?

Solution: Assume advertisement expenditure (y) is the dependent variable and sales (x) is the independent variable. Then, the regression equation for advertisement expenditure on sales is given by

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Given $r = 0.8$, $\sigma_x = 40$, $\sigma = 25$, $\bar{x} = 45{,}000$, $\bar{y} = 30{,}000$. Substituting this value in the above equation, we have

$$(y - 30{,}000) = 0.8\ \frac{25}{40}\ (x - 45{,}000) = 0.5\ (x - 45{,}000)$$

$$y = 30{,}000 + 0.5x - 22{,}500 = 7500 + 0.5x$$

When a sales target is fixed at $x = 80{,}000$, the estimated amount likely to be spent on advertisement would be

$$y = 7500 + 0.5 \times 80{,}000 = 7500 + 40{,}000 = \text{Rs } 47{,}500$$

Example 12.5: You are given the following information about advertising expenditure and sales:

Statistic	Advertisement (x) (Rs in lakh)	Sales (y) (Rs in lakh)
Arithmetic Mean	10	90
Standard Deviation	3	12

Correlation coefficient = 0.8

1. Obtain the two regression equations.
2. Find the likely sales when the advertisement budget is Rs 15 lakh.
3. What should the advertisement budget be if the company wants to attain a sales target of Rs 120 lakh?

Solution:

(a) The Regression equation of x on y is given by

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Given $\bar{x} = 10$, $r = 0.8$, $\sigma_x = 3$, $\sigma_y = 12$, $\bar{y} = 90$. Substituting these values in the above regression equation, we have

$$x - 10 = 0.8\ \frac{3}{12}\ (y - 90) \quad \text{or} \quad x = -8 + 0.2y$$

The regression equation of y on x is given by

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 90 = 0.8\ \frac{12}{3}(x - 10) \quad \text{or} \quad y = 58 + 3.2x$$

(b) Substituting $x = 15$ in the regression equation of y on x. The likely average sales volume would be

$$y = 58 + 3.2\ (15) = 58 + 48 = 106$$

Thus, the likely sales for the advertisement budget of Rs 15 lakh is Rs 106 lakh.

(c) Substituting $y = 120$ in the regression equation of x on y. The likely advertisement budget to attain the desired sales target of Rs 120 lakh would be

$$x = -8 + 0.2\,y = -8 + 0.2\,(120) = 16$$

Hence, the likely advertisement budget of Rs 16 lakh should be sufficient to attain the sales target of Rs 120 lakh.

Example 12.6: In a partially destroyed laboratory record of an analysis of regression data, the following results are only legible:
Variance of $x = 9$
Regression equations: $8x - 10y + 66 = 0$ and $40x - 18y = 214$
Find based on the above information:

1. The mean values of x and y
2. Coefficient of correlation between x and y and
3. The standard deviation of y.

Solution: (a) Since two regression lines always intersect at a point $(\overline{x},\overline{y})$ representing mean values of the variables involved, solving the given regression equations to get the mean values $\overline{xx}$ as shown below:

$$8x - 10y = -66$$
$$40x - 18y = 214$$

Multiplying the first equation by 5 and subtracting it from the second, we have

$$32y = 544 \quad \text{or} \quad y = 17, \text{ i.e. } \overline{y} = 17$$

Substituting the value of y in the first equation, we get

$$8x - 10(17) = -66 \quad \text{or} \quad x = 13, \text{ that is, } \overline{x} = 13$$

(b) To find the correlation coefficient r between x and y, we need to determine the regression coefficients b_{xy} and b_{xy}
Rewriting the given regression equations in such a way that the coefficient of the dependent variable is less than one, at least in one equation.

$$8x - 10y = -66 \text{ or } 10y = 66 + 8x \text{ or } y = \frac{66}{10} + \frac{8}{10}x$$

That is, $b_{yx} = 8/10 = 0.80$

$$40x - 18y = 214 \text{ or } 40x = 214 + 18y \text{ or } x = \frac{214}{40} + \frac{18}{40}y$$

That is, $b_{xy} = 18/40 = 0.45$
Hence, the coefficient of correlation r between x and y is given by

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.45 \times 0.80} = 0.60$$

(c) To determine the standard deviation of y, consider the formula:

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \text{or} \quad \sigma_y = \frac{b_{yx}\,\sigma_x}{r} = \frac{0.80 \times 3}{0.6} = 4$$

Example 12.7: The two regression lines obtained in a correlation analysis of 60 observations are:

$$5x = 6x + 24 \quad \text{and} \quad 1000y = 768x - 3708$$

What is the correlation coefficient, and what is its probable error? Show that the ratio of the coefficient of variability of x to that of y is 5/24. What is the ratio of variances of x and y?
Solution: Rewriting the regression equations

$$5x = 6y + 24 \quad \text{or} \quad x = \frac{6}{5}y + \frac{24}{5}$$

That is, $b_{xy} = 6/5$

$$1000y = 768x - 3708 \quad \text{or} \quad y = \frac{768}{1000}x - \frac{3708}{1000}$$

That is, $b_{yx} = 768/1000$

We know that $b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$ and $b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{768}{1000}$, therefore

$$b_{xy}b_{yx} = r^2 = \frac{6}{5}\times\frac{768}{1000} = 0.9216$$

Hence $r = \sqrt{0.9216} = 0.96$.

The correlation coefficient is positive since both bxy and byx are positive, and hence $r = 0.96$.

$$\text{Probable error of } r = 0.6745\frac{1-r^2}{\sqrt{n}} = 0.6745\frac{1-(0.96)^2}{\sqrt{60}}$$

$$= \frac{0.0528}{7.7459} = 0.0068$$

Solving the given regression equations for x and y, we get $\bar{x} = 6$ and $\bar{y} = 1$ because the regression lines passed through the point $(\bar{x},\bar{y})$.

$$\text{Since } r\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } 0.96\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } \frac{\sigma_x}{\sigma_y} = \frac{6}{5\times 0.96} = \frac{5}{4}$$

$$\text{Also the ratio of the coefficient of variability} = \frac{\sigma_x/\bar{x}}{\sigma_y/\bar{y}} = \frac{\bar{y}}{\bar{x}}\cdot\frac{\sigma_x}{\sigma_y} = \frac{1}{6}\times\frac{5}{4} = \frac{5}{24}.$$

12.2 COMPARISON BETWEEN LINEAR CORRELATION AND REGRESSION:

Aspect	Correlation	Regression
Measurement Level	Interval or ratio scale	Interval or ratio scale
Nature of Variables	Both continuous and linearly related	Both continuous and linearly related
x–y Relationship	x and y are symmetric	y is dependent, x is independent; regression of x on y differs from y on x
Correlation Coefficient	*bxy=byx*	The correlation between x and y is the same as the correlation between y and x
Coefficient of Determination	Explain the common variance of x and y	Proportion of variability of x explained by its least-squares regression on y

12.3 SUMMARY:

Regression is a statistical method used to examine the relationship between a dependent variable and one or more independent variables. It helps predict values and understand trends. Regression analysis offers advantages such as simplicity, interpretability, and practical application. There are various regression models, with linear regression being the most basic. The least squares method estimates the regression line by minimizing the sum of squared errors. Regression coefficients are calculated to represent the relationship, and their properties help assess the model's strength and direction.

12.4 TECHNICAL TERMS:

- **Regression Analysis:** A technique to model and analyze the relationship between variables, especially for prediction.
- **Dependent Variable:** The variable whose value is predicted or explained (often denoted as Y).
- **Independent Variable:** The variable used to predict or explain the dependent variable (often denoted as X).
- **Simple Linear Regression:** A regression model involving one independent and one dependent variable with a linear relationship.
- **Regression Coefficients:** Constants that quantify the relationship between the independent and dependent variables (slope and intercept).

12.5 SELF-ASSESSMENT PROBLEMS:

1. The following calculations have been made for prices of 12 stocks (x) at the Calcutta Stock Exchange on a particular day, along with the sales volume in thousands of shares (y). From these calculations, find the regression equation of the price of stocks on the volume of shares sold.

   $$\Sigma x = 580,\ \Sigma y = 370,\ \Sigma xy = 11494,\ \Sigma x^2 = 41658,\ \Sigma y^2 = 17206$$

2. The following data give the experience of machine operators and their performance ratings, given by the number of good parts turned out per 100 pieces:

Operator	1	2	3	4	5	6	7	8
Experience(x)	16	12	18	4	3	10	5	12
Performance ratings(y)	87	88	89	68	78	80	75	83

   Calculate the regression lines of performance ratings on experience and estimate the probable performance if an operator has 7 years of experience.

3. The following table gives the aptitude test scores and productivity indices of 10 workers selected at random:

Aptitude scores(x)	60	62	65	70	72	48	53	73	65	82
Productivity index(y)	68	60	62	80	85	40	52	62	60	81

   Calculate the two regression equations and estimate (a) the productivity index of a worker whose test score is 92 and (b) the test score of a worker whose productivity index is 75.

4. A company wants to assess the impact of R&D expenditure (Rs in 1000s) on its annual profit (Rs in 1000s). The following table presents the information for the last eight years:

Year	*R&D Expenditure*	*Annual Profit*
1991	9	45
1992	7	42
1993	5	41
1994	10	60

1995	4	30
1996	5	34
1997	3	25
1998	2	20

Estimate the regression equation and predict the annual profit for the year 2002 for an allocated sum of Rs 1,00,000 as R&D expenditure.

5. The personnel manager of an electronics manufacturing company devises a manual test for job applicants to predict their production rating in the assembly department. To do this, he selects a random sample of 10 applicants. They are given the test and later assigned a production rating. The results are as follows:

Worker	A	B	C	D	E	F	G	H	I	J
Test Score	53	36	88	84	86	64	45	48	39	69
Rating	45	43	89	79	84	66	49	48	43	76

Fit a linear least squares regression equation of production rating on test score.

6. Two random variables have the regression equations:
$3x + 2y - 26 = 0$ and $6x + y - 31 = 0$
   a) Find the mean values of x and y and the correlation coefficient between x and y.
   b) If the variance of x is 25, then find the standard deviation of y from the data.

7. Explain the concept of regression and point out its usefulness in dealing with business problems.

8. Distinguish between correlation and regression. Also, point out the properties of the regression coefficient.

12.6 REFERENCES :

1. Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.
2. Levin, R.I., & Rubin, D.S. (2017). *Statistics for Management* (7th ed.). Pearson Education.
3. Sharma, J.K. (2020). *Business Statistics*. Pearson Education India.

Dr. G. Malathi

LESSON- 13
TIME SERIES ANALYSIS-1

OBJECTIVES:

After studying this unit, you should be able to:

- Clarify the definition of time series,
- Recognize the significance of time series in short-term forecasting,
- Describe the various components that make up a time series, and
- Calculate trend values using various techniques.

STRUCTURE:

13.1 INTRODUCTION:

In the previous units, you have learnt statistical treatment of data collected for research work. The nature of data varied from case to case. You have come across quantitative data for a group of respondents collected with a view to understanding one or more parameters of that group, such as investment, profit, consumption, weight etc. But when a nation, state, an institution or a business unit etc., intend to study the behaviour of some element, such as price of a product, exports of a product, investment, sales, profit etc., as they have behaved over a period of time, the information shall have to be collected for a fairly long period, usually at equal time intervals. Thus, a set of any quantitative data collected and arranged on the basis of time is called 'Time Series'.

Depending on the research objective, the unit of time may be a decade, a year, a month, or a week etc. Typical time series are the sales of a firm in successive years, monthly production figures of a cement mill, daily closing price of shares in Bombay stock market, hourly temperature of a patient.

Usually, the quantitative data of the variable under study are denoted by
y_1, y_2,-------------------------------------y_n and the corresponding time units are denoted by
t_1, t_2---t_n. The variable
'y' shall have variations, as you will see ups and downs in the values. These changes account

for the behaviour of that variable.

Instantly it comes to our mind that 'time' is responsible for these changes, but this is not true. Because, the time (t) is not the cause and the changes in the variable (y) are not the effect. The only fact, therefore, which we must understand is that there are a number of causes which affect the variable and have operated on it during a given time period. Hence, time becomes only the basis for data analysis.

Forecasting any event helps in the process of decision making. Forecasting is possible if we are able to understand the past behaviour of that particular activity. For understanding the past behaviour, a researcher needs not only the past data but also a detailed analysis of the same.

Thus, in this unit we will discuss the need for analysis of time series, fluctuations of time series which account for changes in the series over a period of time, and measurement of trend for forecasting. A time series is an arrangement of statistical data in a chronological order, *i.e.,* in accordance with its time of occurrence. It reflects the dynamic pace of movements of a phenomenon over a period of time. Most of the series relating to Economics, Business and Commerce, *e.g.,* the series relating to prices, production and consumption of various commodities; agricultural and industrial production, national income and foreign exchange reserves ; investment, sales and profits of business houses ; bank deposits and bank clearings, prices and dividends of shares in a stock exchange market, etc., are all time series spread over a long period of time. Accordingly, time series have an important and significant place in Business and Economics, and basically most of the statistical techniques for the analysis of time series data have been developed by economists. However, these techniques can also be applied for the study of behaviour of any phenomenon collected chronologically over a period of time in any discipline relating to natural and social sciences, though not directly related to economics or business.

13.1.1 Definition and Utility of Time Series Analysis

"*A time series may be defined as a collection of readings belonging to different time periods, of some economic variable or composite of variables*".

Mathematically, a time series is defined by the functional relationship

$$y = f(t)$$

where y is the value of the phenomenon (or variable) under consideration at time t. For example,

(i) the population (y) of a country or a place in different years (t),
(ii) the number of births and deaths (y) in different months (t) of the year,
(iii) the sale (y) of a departmental store in different months (t) of the year,
(iv) the temperature (y) of a place on different days (t) of the week, and so on constitute time series. Thus, if the values of a phenomenon or variable at times $t_1, t_2, \ldots, t_n$ are $y_1, y_2, \ldots, y_n$ respectively, then the series

t :	t_1	t_2	t_3	…	
t_n y :		y_1	y_2	y_3	
…	y_n				

constitutes a time series. Thus, a time series invariably gives a bivariate distribution, one of the two variables being time (t) and the other being the value (y) of the phenomenon at different points of time. The values of t may be given yearly, monthly, weekly, daily or even hourly, usually but not always at equal intervals of time. The graph of a time series, known as *Histogram*, is obtained on plotting the data on a graph paper taking the independent variable

t along the x-axis and the dependent variable y along the y-axis.
The analysis of time series is of great utility not only to research workers but also to economists, businessmen and scientists etc., for the following reasons:

1) It helps in understanding past behaviour of the variables under study.
2) It facilitates in forcasting the future behaviour with the help of the changes that have taken place in the past.
3) It helps in planning future course of action.
4) It helps in knowing current accomplishment.
5) It is helpful to make comparisons between different time series and significant conclusions drawn therefrom.

Thus we can say at the need for time series analysis arises in research because:

- we want to understand the behaviour of the variables under study,
- we want to know the expected quantitative changes in the variable under study
- we want to estimate the effect of various causes in quantitative terms.
- In a nutshell, the time series analysis is not only useful for researchers, business research institutions, but also for Governments for devising appropriate future growth strategies.

13.2 COMPONENTS OF A TIME SERIES:

If the values of a phenomenon are observed at different periods of time, the values so obtained will show appreciable variations or changes. These fluctuations are due to the fact that the value of the phenomenon is affected not by a single factor but due to the cumulative effect of a multiplicity of factors pulling it up and down. However, if the various forces were in a state of equilibrium, then the time series will remain constant. For example, the sales (y) of a product are influenced by (*i*) advertisement expenditure, (*ii*) the price of the product, (*iii*) the income of the people, (*iv*) other competitive products in the market, (*v*) tastes, fashions, habits and customs of the people and so on. Similarly, the price of a particular product depends on its demand, various competitive products in the market, raw materials, transportation expenses, investment, and so on. The various forces affecting the values of a phenomenon in a time series may be broadly classified into the following four categories, commonly known as the *components of a time series*, some or all of which are present (in a given time series) in varying degrees.

(a) Secular Trend or Long-term Movement (T).
(b) Periodic Movements or Short-term Fluctuations :
(i) Seasonal Variations (S), (*ii*) Cyclical Variations (C).
(c) Random or Irregular Variations (R or I).

The value (y) of a phenomenon observed at any time (t) is the net effect of the interaction of above components. We shall explain these components briefly in the following sections.

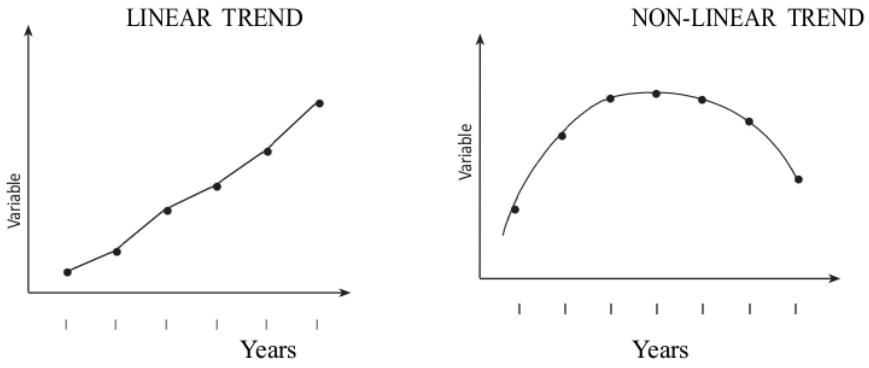
Secular Trend. The general tendency of the time series data to increase or decrease or stagnate during a *long* period of time is called the *secular trend* or *simple trend.* This phenomenon is usually observed in most of the series relating to Economics and Business, *e.g.,* an upward tendency is usually observed in time series relating to population, production and sales of products, prices, incomes, money in circulation, etc., while a downward tendency is noticed in the time series relating to deaths, epidemics, etc., due to advancement in medical technology, improved medical facilities, better sanitation, diet, etc.

According to Simpson and Kafka :

"*Trend, also called secular or long-term trend, is the basic tendency of a series…to grow or decline over a period of time. The concept of trend does not include short-range oscillations, but rather the steady movement over a long time.*"

1. The term '*long period of time*' is a relative term and cannot be defined exactly. It would very much depend on the nature of the data. In certain phenomenon, a period as small as few hours may be sufficiently long, while in others even a period as long as 3- 4 years may not be sufficient. For example, to have an idea about the production of a particular product (agricultural or industrial production), an increase over the past 20 or 30 months will not reflect a secular change for which we must have data for 7-8 years. In such a phenomenon, the values for short period (2-3 years) are unduly affected by cyclic variation (discussed later) and will into reveal the true trend. In order to have true picture of the trend, the time series values must be examined over a period covering at least two or three complete cycles.On the other hand, if we count the number of bacterial population (living organisms) of a culture subjected to strong germicide every 20 seconds for 1 hour, then the set of 180 readings showing a general pattern would be termed as secular movement.

2. *Linear and Non-Linear* (*Curvi-Linear*) *Trend.* If the time series values plotted on graph cluster more or less round a straight line, the trend exhibited by the time series is termed as *Linear* otherwise *Non-Linear* (*Curvi-Linear*)—See Figures 11·1 and 11·2. In a straight line trend, the time series values increase or decrease more or less by a constant absolute amount, *i.e.,* the rate of growth (or decline) is constant. Although, in practice, linear trend is commonly used, it is rarely observed in economic and business data. In an economic and business phenomenon, the rate of growth or decline is not of constant nature throughout but varies considerably in different sectors of time. Usually, in the beginning, the growth is slow, then rapid which is further accelerated for quite some time, after which it becomes stationary or stable for some period and finally retards slowly.



3. It is not necessary that all the series must exhibit a rising or declining trend. Certain phenomena may give rise to time series whose values fluctuate round a constant reading which does not change with time, *e.g.,* the series relating to temperature or barometric readings (pressure) of a particular place.

4. Uses of Trend. (*i*) The study of the data over a long period of time enables us to have a general idea about the pattern of the behaviour of the phenomenon under consideration. This

helps in business forecasting and planning future operations. For example, if the time series data for a particular phenomenon exhibits a trend in a particular direction, then under the assumption that the same pattern will continue in the near future, an assumption which is quite reasonable unless there are some fundamental and drastic changes in the forces affecting the phenomenon—we can forecast the values of the phenomenon for future also.

The accuracy of the trend curve or trend equation or the estimates obtained from them will depend on the reliability of the type of trend fitted to the given data. (For details, see Measurement of Trend - Least Square Method.) The trend values are of paramount importance to a businessman in providing him the rough estimates of the values of the phenomenon in the near future. For instance, an idea about the approximate sales or demand for a product is extremely useful to a businessman in planning future operations and formulating policies regarding inventory, production, etc.

(ii) By isolating trend values from the given time series, (By dividing the given time series values by the trend values or subtracting trend values from the given time series values we can study the short-term and irregular movements.

(iii) Trend analysis enables us to compare two or more time series over different periods of time and draw important conclusions about them.

Short-Term Variations. In addition to the long-term movements there are inherent in most of the time series, a number of forces which repeat themselves periodically or almost periodically over a period of time and thus prevent the smooth flow of the values of the series in a particular direction. Such forces give rise to the so-called *short-term variations* which may be classified into the following two categories :
(i) Sesonal Variations (*S*), and (*ii*) Cyclical Variations (*C*).

Seasonal Variations (*S*). These variations in a time series are due to the rhythmic forces which operate in a regular and periodic manner over a span of less than a year, *i.e.,* during a period of 12 months and have the same or almost same pattern year after year. Thus, seasonal variations in a time series will be there if the data are recorded quarterly (every three months), monthly, weekly, daily, hourly, and so on. Although in each of the above cases, the amplitudes of the seasonal variations are different, all of them have the same period, *viz.,* 1 year. Thus *in a time series data where only annual figures are given, there are no seasonal variations.* Most of economic time series are influenced by seasonal swings, *e.g.,* prices, production and consumption of commodities ; sales and profits in a departmental store ; bank clearings and bank deposits, etc., are all affected by seasonal variations. The seasonal variations may be attributed to the following two causes :

(i) *Those resulting from natural forces.* As the name suggests, the various seasons or weather conditions and climatic changes play an important role in seasonal movements. For instance, the sales of umbrella pick up very fast in rainy season ; the demand for electric fans goes up in summer season ; the sales of ice and ice-cream increase very much in summer ; the sales of woollens go up in winter - all being affected by natural forces, *viz*., weather or seasons. Likewise, the production of certain commodities such as sugar, rice, pulses, eggs, etc., depends on seasons. Similarly, the prices of agricultural commodities always go down at the time of harvest and then pick up gradually.

(ii) *Those resulting from man-made conventions.* These variations in a time series within a

period of 12 months are due to habits, fashions, customs and conventions of the people in the society. For instance, the sales of jewellery and ornaments go up in marriages ; the sales and profits in departmental stores go up considerably during marriages, and festivals like Deepawali, Dushehra (Durga Pooja), Christmas, etc. Such variations operate in a regular spasmodic manner and recur year after year.

The main objective of the measurement of seasonal variations is to isolate them from the trend and study their effects. A study of the seasonal patterns is extremely useful to businessmen, producers, sales- managers, etc., in planning future operations and in formulation of policy decisions regarding purchase, production, inventory control, personnel requirements, selling and advertising programmes.

Cyclical Variations (C). The oscillatory movements in a time series with period of oscillation greater than one year are termed as *cyclical variations*. These variations in a time series are due to ups and downs recurring after a period greater than one year. The cyclical fluctuations, though more or less regular, are not necessarily uniformly periodic, *i.e.,* they may or may not follow exactly similar patterns after equal intervals of time. One complete period which normally lasts from 7 to 9 years is termed as a '*cycle*'. These oscillatory movements in any business activity are the outcome of the so-called '*Business Cycles*' *which are the four-phased cycles comprising prosperity (boom), recession, depression and recovery from time to time.* These booms and depressions in any business activity follow each other with steady regularity and the complete cycle from the peak of one boom to the peak of next boom usually lasts from 7 to 9 years. Most of the economic and business series, *e.g.,* series relating to production, prices, wages, investments, etc., are affected by cyclical upswings and downswings.

The study of cyclical variations is of great importance to business executives in the formulation of policies aimed at stabilising the level of business activity. A knowledge of the cyclic component enables a businessman to have an idea about the periodicity of the booms and depressions and accordingly he can take timely steps for maintaining stable market for his product.

Random or Irregular Variations. Mixed up with cyclical and seasonal variations, there is inherent in every time series another factor called *random or irregular variations*. These fluctuations are purely random and are the result of such unforeseen and unpredictable forces which operate in absolutely erratic and irregular manner. Such variations do not exhibit any definite pattern and there is no regular period or time of their occurrence, hence they are named irregular variations. These powerful variations are usually caused by numerous non-recurring factors like floods, famines, wars, earthquakes, strikes and lockouts, epidemics, revolution, etc., which behave in a very erratic and unpredictable manner. Normally, they are short-term variations but sometimes their effect is so intense that they may give rise to new cyclical or other movements. Irregular variations are also known as *episodic* fluctuations and include all types of variations in a time series data which are not accounted for by trend, seasonal and cyclical variations.

Because of their absolutely random character, it is not possible to isolate such variations and study them exclusively nor we can forecast or estimate them precisely. The best that can be done about such variations is to obtain their rough estimates (from past experience) and accordingly make provisions for such abnormalities during normal times in business.

13.3 MEASUREMENT OF TREND:

The following are the four methods which are generally used for the study and measurement of the trend component in a time series.

(i) *Graphic* (or *Free-hand* Curve Fitting) Method.
(ii) Method of *Semi-Averages.*
(iii) Method of *Curve Fitting* by the *Principle of Least Squares.*
(iv) Method of *Moving Averages*.

13.3.1. Graphic or Free Hand Curve Fitting Method

This is the simplest and the most flexible method of estimating the secular trend and consists in first obtaining a historigram by plotting the time series values on a graph paper and then drawing a free-hand smooth curve through these points so that it accurately reflects the long-term tendency of the data. The smoothing of the curve eliminates the other components, *viz*., seasonal, cyclical and random variations. In order to obtain proper trend line or curve, the following points may be borne in mind :

(i) It should be smooth.
(ii) The number of points above the trend curve/line should be more or less equal to the number of points below it.
(iii) The sum of the vertical deviations of the given points above the trend line should be approximately equal to the sum of vertical deviations of the points below the trend line so that the total positive deviations are more or less balanced against total negative deviations.
(iv) The sum of the squares of the vertical deviations of the given points from the trend line/curve is minimum possible.
[The points (*iii*) and (*iv*) conform to the principle of average (Arithmetic Mean) because the algebraic sum of the deviations of the given observations from their arithmetic mean is zero and the sum of the squared deviations is minimum when taken about mean.]
(v) If the cycles are present in the data then the trend line should be so drawn that :
(a) It has equal number of cycles above and below it.
(b) It bisects the cycles so that the areas of the cycles above and below the trend line are approximately same.
(vi) The minor short-term fluctuations or abrupt and sudden variations may be ignored.

Merits.

(*i*) It is very simple and time-saving method and does not require any mathematical calculations.
(*ii*) It is a very flexible method in the sense that it can be used to describe all types of trend – linear as well as non-linear.

Demerits.

(*i*) The strongest objection to this method is that it is highly subjective in nature. The trend curve so obtained will very much depend on the personal bias and judgement of the investigator handling the data and consequently different persons will obtain different trend curves for the same set of data. Thus, a proper and judicious use of this method requires great skill and expertise on the part of the investigator and this very much restricts the popularity and utility of this method. This method, though simple and flexible, is seldom used in practice because of the inherent bias of the investigator.
(ii) It does not help to measure trend.

(iii) Because of the subjective nature of the free-hand trend curve, it will be dangerous to use it for forecasting or making predictions.

13.3.2. Method of Semi-Averages

As compared with graphic method, this method has more objective approach. In this method, the whole time series data is classified into two equal parts *w.r.t.* time. For example, if we are given the time series values for 10 years from 1985 to 1994 then the two equal parts will be the data corresponding to periods 1985 to 1989 and 1990 to 1994.

However, in case of odd number of years, the two equal parts are obtained on omitting the value for the middle period. Thus, for example, for the data for 9 years from 1990 to 1998, the two parts will be the data for years 1990 to 1993 and 1995 to 1998, the value for the middle year, *viz*., 1994 being omitted. Having divided the given series into two equal parts, we next compute the arithmetic mean of time-series values for each half separately.

These means are called *semi-averages.* Then these semi-averages are plotted as points against the middle point of the respective time periods covered by each part. The line joining these points gives the straight line trend fitting the given data.

As an illustration, for the time series data for 1985 to 1994, we have :

	Part I	*Part II*
Period :	1985 to 1989	1990 to 1994
Semi-Average :	$\bar{x}_1 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$	$\bar{x}_2 = \frac{y_6 + y_7 + \ldots + y_{10}}{5}$
Middle of time period :	1987	1992

$\bar{x}_1$ is plotted against 1987 and $\bar{x}_2$ is plotted against 1992. The trend line is obtained on joining the points so obtained, *viz.,* the points (1987, $\bar{x}_1$) and (1992, $\bar{x}_2$) by a straight line. In the above case, the two parts consisted of an odd number of years, *viz.,* 5 and hence the middle time period is computed easily. However, if the two halves consist of even numbers of years as in the next case given above; *viz*., the years 1990 to 1993 and 1995 to 1998, the centring of average time period is slightly difficult. In this case $\bar{x}_1$ (the mean of the values for the years 1990 to 1993) will be plotted against the mean of the two middle years of the period 1990 to 1993, *viz*., the mean of the years 1991 and 1992. Similarly, $\bar{x}_2$ will be plotted against the mean of the years 1996 and 1997.

Merits.

(*i*) An obvious advantage of this method is its objectivity in the sense that it does not depend on personal judgement and everyone who uses this method gets the same trend line and hence the same trend values.

(*ii*) It is easy to understand and apply as compared with the moving average or the least square methods of measuring trend.

(*iii*) The line can be extended both ways to obtain future or past estimates.

Limitations.

(*i*) This method assumes the presence of linear trend (in the time series values) which may not exist.

(*ii*) The use of arithmetic mean (for obtaining semi-averages) may also be questioned

because of its limitations.

Accordingly, the trend values obtained by this method and the predicted values for future are not precise and reliable.

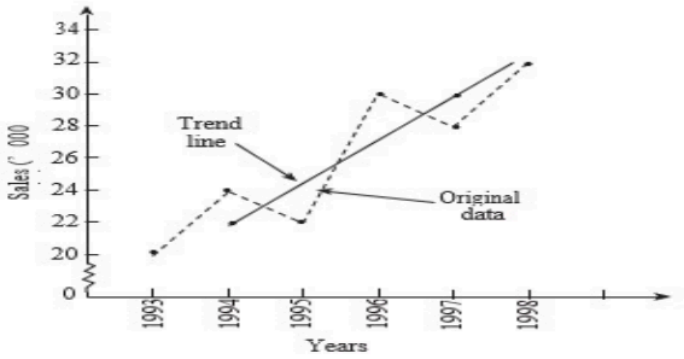
Problem *Apply the method of semi-averages for determining trend of the following data and estimate the value for 2000* :

Years	:	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*
Sales (*thousand units*)	:	*20*	*24*	*22*	*30*	*28*	*32*

If the actual figure of sales for 2000 is 35,000 units, how do you account for the difference between the figures you obtain and the actual figures given to you ?+

Year	*Sales* (*thousand units*)	*3-Yearly Semi-Totals*	*Semi-Average* (*A.M.*)
1993	20		
1994	24 →	66	$\frac{66}{3} = 22$
1995	22		
1996	30		
1997	28 →	90	$\frac{90}{3} = 30$
1998	32		

Solution. Here $n = 6$ (even), and hence the two parts will be 1993 to 1995 and 1996 to 1998. Here the semi-average 22 is to be plotted against the mid-year of first part, *i.e.,* 1994 and the semi-average 30 is to be plotted against the mid- year of second part, *viz*., 1997. The trend line is shown in the Fig. 11·3.



. COMPUTATION OF TREND VALUES

Year	*Trend Values* (*'000 units*)	*Year*	*Trend Values* (*'000 units*)
1993	22 – 2·667 = 19·333	1997	30
1994	22	1998	30 + 2·667 = 32·667
1995	22 + 2·667 = 24·667	1999	32·667 + 2·667 = 35·334
1996	24·667 + 2·667 = 27·334	2000	35·334 + 2·667 = 38·001

Thus the estimated (trend) value for sales in 2000 is 38,001 units. This trend value differs

from the given value of 35,000 units because it has been obtained under the assumption that there is a linear relationship between the given time series values which in this case (as is obvious from the graph of the original data) is not true. Moreover, in computing the trend value, the effects of seasonal, cyclical and irregular variations have been completely ignored while the observed values are affected by these factors.

PROBLEM . *From the following series of annual data, find the trend line by the method of semi- averages. Also estimate the value for 1999.*

Year	:	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*
Actual Value	:	*170*	*231*	*261*	*267*	*278*	*302*	*299*	*298*	*340*

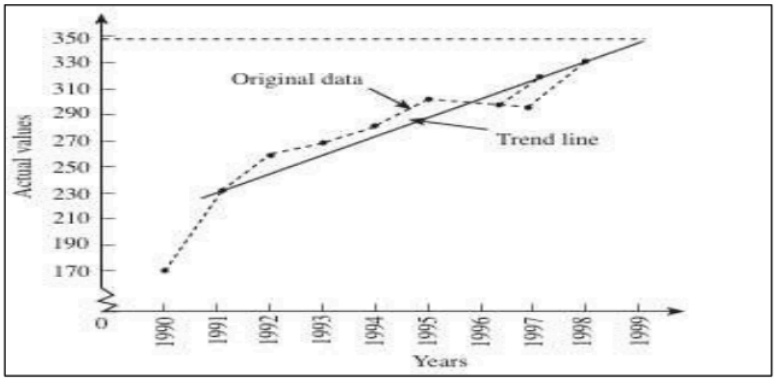
Solution. Here the number of years is 9, *i.e.,* odd. The two middle parts will be 1990 to 1993 and 1995 to 1998, the value for middle year, *viz.,* 1994 being ignored.

Year	*Actual Value*	*4 Yearly Semi-Totals*	*Semi-Average*
1990	170		
1991	231		
1992	261 →	929	$\frac{929}{4} = 232{\cdot}25 \simeq 232$
1993	267		
1994	278		
1995	302		
1996	299		
1997	298 →	1239	$\frac{1239}{4} = 309{\cdot}75 \simeq 310$
1998	340		

The value 232 is plotted against the middle of the years 1991 and 1992 and the value 310 is plotted against the middle of the years 1996 and 1997. The trend line graph is shown in the below figure
From the graph we see that the estimated (trend) value for 1999 is 348.

Aliter. *Trend Value for 1999* : From the calculations in the above table we observe that the increment in the actual value from middle of 1991- 92 to the middle of 1996-97, *i.e.,* for 5 years is 310 – 232 = 78. Hence the yearly increment is 78/5. We also find that the average trend value for middle of 1996-97 is 310. Hence the trend value for 1999 is given by

$$310 + (5/2) \times (78/5) = 310 + 39 = 349.$$



This value differs from the graph value of 348 obtained from the trend line because of the reason and also because we have obtained the calculations by rounding the decimals.

13.4 SUMMARY:

This unit has introduced you to the concept of time series and its analysis with a view to making more accurate and reliable forcasts for the future.
A set of quantitative data arranged on the basis of TIME are referred to as 'Time Series'. The analysis of time series is done to understand the dynamic conditions for achieving the short-term and long-term goals of institution(s). With the help of the techniques of time series analysis the future pattern can be predicted on the basis of past trends.

The quantitative values of the variable under study are denoted by y_1, y_2, y_3..... and the corresponding time units are denoted as x_1, x_2, x_3.....The variable 'y' shall have variations, you will see ups and downs in the values. There are a number of causes during a given time period which affect the variable. Therefore, time becomes the basis of analysis. Time is not the cause and the changes in the values of the variable are not the effect.The causes which affect the variable gradually and permanently are termed as Long-term causes. The causes which affect the variable only for the time being are termed as Short-term causes. The time series are usually the result of the effects of one or more of the four components. These are trend variations (T), seasonal variations (S), Cyclical variations (C) and Irregular variations (I).

13.5 KEY WORDS:

1. **Time Series :** is the data on any variable accumulated at regular time intervals.
2. **Secular Trend :** A type of variation in a time series, the long-term tendency of a time series to grow or decline over a period of time.
3. **Seasonal Variation :** Patterns of change in a time series within a year and the same changes tend to be repeated from year to year.
4. **Cyclical Variations :** A type of variation in a time series, in which the values of variables vary up and down around the secular trend line.
5. **Irregular Variations :** A type of element of a time series, refers to such variations in business activity which do not repeat according to a definite pattern and the values of variables are completely unpredictable.

13.6 SELF ASSESSMENT QUESTIONS:

1) What is time series? Why do we analyse a time series?
2) Explain briefly the components of time series.
3) Explain briefly the additive and multiplicative models of time series. Which of these models is more commonly used and why?
4) From the following data, obtain the trend line by Freehand Method for further analysis.

Years	1996	1997	1998	1999	2000	2001	2002	2003
'y'	24	28	38	33	49	50	66	68

13.7 SUGGESTED READINGS;

- Mentgomery, D.C. and L.A. Johnson, 1996, '*Forecasting and Time Series Analysis*' McGraw Hill : New York.
- Chandan, J.S., 2001, *Statistics for Business and Economics*, Vikas Publishing House Pvt. Ltd., New Delhi.
- Gupta, S.P. and H.P. Gupta, 2001, *Business Statistics*, S. Chand, New Delhi.

Dr. Naga Nirmala Rani

LESSON-14

TIME SERIES ANALYSIS- 2

OBJECTIVES:

After studying this unit, you should be able to:

- Calculate the trend values using various techniques.
- Determine the trend values based on the least squares principle.
- Assess the trend using the moving averages method.

STRUCTURE:

14.1 METHOD OF CURVE FITTING BY THE PRINCIPLE OF LEAST SQUARES:

The principle of least squares serves as a mathematical tool that enables us to achieve an objective alignment with the trend of a specified time series. A significant portion of data associated with economic and business time series adheres to specific patterns of growth or decline. In these circumstances, employing analytical trend fitting becomes a more dependable method for forecasting and making predictions. This approach is applicable for fitting both linear and non-linear trends.

Fitting of Linear Trend. Let the straight line trend between the given time-series values (y) and time(t) be given by the equation :

$$y = a + bt$$

Then for any given time 't', the estimated value y_e of y as given by this equation is :

$$y_e = a + bt$$

the principle of least squares consists in estimating the values of a and b in the above equation so that the sum of the squares of errors of estimate

$$E = \text{å}(y - y_e)^2 = \text{å}(y - a - bt)^2,$$

is minimum, the summation being taken over given values of the time series.
which, on simplification, gives the *normal equations* or *least square equations* for estimating a and b as

$$\text{å}y = na + b\text{å}\,t \qquad \text{and} \qquad \text{å}ty = a\text{å}t + b\text{å}t^2,$$

where is the number of time series pairs (t, y). It may be seen that equation is obtained on taking sum of both sides in equation .the resultant equation is obtained on multiplying equation by t and then summing both sides over the given values of the series.

Solving both the equations for a and b and substituting these values in we finally get the equation of the straight line trend.

2.The straight line trend implies that irrespective of the seasonal and cyclical swings and irregular fluctuations, the trend values increase or decrease by a constant absolute amount 'b' per unit of time. Thus, if we are given the yearly figures for a time series, then the coefficient 'b' in the line (11·9), which is nothing but the *slope* of the trend line [*c.f.* equation of a line in the form : $y = mx + c$], gives the *annual rate of growth*. Hence, the *linear trend values form a series in arithmetic progression, the common difference being* 'b', *the slope of the trend line*.

After obtaining the trend line by the principle of least squares, the trend values for different years can be obtained on substituting the values of time t in the trend equation. However, from practical point of view, a much more convenient method of obtaining the trend values of different years is to compute the trend value for the first year from the equation of the trend line and then add the value of 'b' to it successively (because the trend values form a series in A.P. with common difference 'b').

Fitting a Second Degree (Parabolic) Trend. Let the second degree parabolic trend be given by the equation :

$y = a + bt + ct^2$ Then for any given value of t, the trend value is given by :

$$y_e = a + bt + ct^2$$

Thus, if y_e is the trend value corresponding to an observed value y, then according to the principle of least squares we have to obtain the values of a, b and c so that

$$E = \text{å}\,(y - y_e)^2 = \text{å}(y - a - bt - ct^2)^2$$

is minimum for variations in a, b and c. Thus, the normal or least square equations for estimating a, b and c are given by :

$$\left.\begin{aligned} \frac{\partial E}{\partial a} &= 0 = -2\sum (y - a - bt - ct^2) \\ \frac{\partial E}{\partial b} &= 0 = -2\sum t\,(y - a - bt - ct^2) \\ \frac{\partial E}{\partial c} &= 0 = -2\sum t^2\,(y - a - bt - ct^2) \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \sum y &= na + b\sum t + c\sum t^2 \\ \sum ty &= a\sum t + b\sum t^2 + c\sum t^3 \\ \sum t^2 y &= a\sum t^2 + b\sum t^3 + c\sum t^4 \end{aligned}\right\}$$

e summation being taken over the values of the time series.

For given time series, the values åy, å ty, å t^2y, å t, å t^2, å t^3 and å t^4 can be calculated and equations (11·15) can be solved for a, b and c. With these values of a, b, c, the parabolic curve (11·14) is the trend curve of best fit.

Merits and Limitations of Trend Fitting by Principle of Least Squares

Merits. The method of least squares is the most popular and widely used method of fitting mathematical functions to a given set of observations. It has the following advantages :

(i) Because of its analytical or mathematical character, this method completely eliminates the element of subjective judgement or personal bias on the part of the investigator.
(ii) Unlike the method of moving averages (discussed in § 11·5·6), this method enables us to compute the trend values for all the given time periods in the series.
(iii) The trend equation can be used to estimate or predict the values of the variable for any period t in future or even in the intermediate periods of the given series and the forecasted values are also quite reliable.
(iv) The curve fitting by the principle of least squares is the *only* technique which enables us to obtain the rate of growth per annum, for yearly data, if linear trend is fitted. If we fit the linear trend $y = a + bx$, where x is obtained from t by change of origin such that $\Sigma x = 0$, then for the yearly data, the annual rate of growth is b or $2b$ according as the number of years is odd or even respectively.

Demerits. (*i*) The most serious limitation of the method is the determination of the type of the trend curve to be fitted, *viz.,* whether we should fit a linear or a parabolic trend or some other more complicated trend curve. [This is discussed in detail in § 11·5·5.] Assumptions about the type of trend to be fitted might introduce some bias.
(i) The addition of even a single new observation necessitates all the calculations to be done afresh which is not so in the case of moving average method.
(ii) This method requires more calculations and is quite tedious and time consuming as compared with other methods. It is rather difficult for a non-mathematical person (layman) to understand and use.
(iii) Future predictions or forecasts based on this method are based only on the long-term variations, *i.e.,* trend and completely ignore the cyclical, seasonal and irregular fluctuations.
(iv) It cannot be used to fit growth curves (Modified exponential curve, Gompertz curve and Logistic curve) to which most of the economic and business time series conform. The discussion, however, is beyond the scope of the book

We shall now discuss some numerical examples to illustrate the technique of curve fitting by the principle of least squares.

Problem *Fit a linear trend to the following data by the least squares method. Verify that* $\Sigma(y - y_e) = 0$, *where* y_e *is the corresponding trend value of* y.

Year :	*1990*	*1992*	*1994*	*1996*	*1998*
Production (in '000 units) :	*18*	*21*	*23*	*27*	*16*

Also estimate the production for the year 1999.[*Delhi Univ. B.Com.* (*Pass*), *1999*]

Solution. Here $n = 5$ *i.e.,* odd. Hence, we shift the origin to the middle of the time period *viz.*, the year 1994.

Let $x = t - 1994$ …(*i*)

Let the trend line of y (production) on x be :

$y = a + bx$ (Origin 1994) …(*ii*)

COMPUTATION OF STRAIGHT LINE TREND

Year (*t*)	*Production ('000 units)*(*y*)	$x = t - 1994$	x^2	xy	*Trend Values ('000 units)* (y_e) = *21 + 0·1x*	$y - y_e$ *('000 units)*
1990	18	– 4	16	–	21 – 0·4 =	–2·6

				72	20·6	
1992	21	–2	4	– 42	21 – 0·2 = 20·8	0·2
1994	23	0	0	0	21·0	2·0
1996	27	2	4	54	21 + 0·2 = 21·2	5·8
1998	16	4	16	64	20 + 0·4 = 21·4	–5·4
	$\Sigma y = 105$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma xy = 4$		$\Sigma(y - y_e) = 0$

The normal equations for estimating a and b in (*ii*) are

$\therefore \Sigma y = na + b\Sigma x$ and $\Sigma xy = a\Sigma x + b\Sigma x^2$

$\Rightarrow$ $105 = 5a + b \times 0$ $\Rightarrow$ $4 = a \times 0 + b \times 40$

$\Rightarrow$ $a = 105/5 = 21$ $\Rightarrow$ $b = 4/40 = 0{\cdot}1$

Substituting in (*ii*), the straight line trend equation is given by : $y = 21 + 0{\cdot}1x$, (Origin : 1994) …(*iii*) [x units = 1 year and y = Production (in '000 units).]

Putting $x = -4, -2, 0, 2$ and 4 in (*iii*), we obtain the trend values (y_e) for the years 1990, 1992,…, 1998 respectively, as given in the last but one column of the Table 11·3.

The difference $(y - y_e)$ is calculated in the last column of the table. We have :

$\Sigma(y - y_e) = -2{\cdot}6 + 0{\cdot}2 + 2{\cdot}0 + 5{\cdot}8 - 5{\cdot}4 = 8 - 8 = 0$, as required.

Estimated Production for 1999. Taking $t = 1999$ in (*i*), we get $x = 1999 - 1994 = 5$.

Substituting $x = 5$ in (*iii*), the estimated production for 1999 is given by :

$(y_e)_{1999} = 21 + 0{\cdot}1 \times 5 = 21 + 0{\cdot}5 = 21{\cdot}5$ thousand units.

Problem. *Below are given the figures of production (in thousand tons) of a sugar factory* :

Year :	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*
Production :	*77*	*88*	*94*	*85*	*91*	*98*	*90*

(i) *Fit a straight line by the method of 'least squares' and show the trend values.*

(ii) *What is the monthly increase in production ?*

(iii) *Eliminate the trend by using both additive and multiplicative models.*

Solution.

COMPUTATION OF STRAIGHT LINE TREND

Year	*Production (in '000 tons)*	$x = t - 2002$	xy	x^2	*Trend Values (in '000 tons)*
					$y_e = 89 + 2x$
1999	77	–3	–231	9	83
2000	88	–2	–176	4	85
2001	94	–1	– 94	1	87
2002	85	0	0	0	89
2003	91	1	91	1	91
2004	98	2	196	4	93

2005	90	3	270	9	95
Total	$\Sigma y = 623$	$\Sigma x = 0$	$\Sigma xy = 56$	$\Sigma x^2 = 28$	$\Sigma y_e = 623$

(i) Let the straight line trend of y on x be given by :

$$y = a + bx \qquad \ldots(*)$$

where the origin is July 2002 and x unit = 1 year. The normal equations for estimating a and b in (*) are :

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

$$\therefore \quad a = \frac{\Sigma y}{n} = \frac{623}{7} = 89 \quad \text{and} \quad b = \frac{\Sigma xy}{\Sigma x^2} = \frac{56}{28} = 2 \qquad [\because \ \Sigma x = 0]$$

Hence, the straight line trend is given by the equation :

$$y = 89 + 2x \quad \text{(Origin : 2002)} \qquad \ldots(**)$$

[x units = 1 year and y = Annual production of sugar (in '000 tons)]

Putting $x = -3, -2, -1, 0, 1, 2, 3$ in (**), we get the trend values for the years 1999 to 2005 respectively and are shown in the last column of the Table 11·4. It may be checked that $\Sigma y = \Sigma y_e$, as required by the principle of least squares.

(ii) From (*) it is obvious that the trend values increase by a constant amount 'b' units every year. Thus, the yearly increase in production is 'b' units, *i.e.,* 2 × 1000 = 2000 tons.

Hence, the monthly increase in production $= \frac{2000}{12} = 166{\cdot}67$ tons.

(iii) Assuming multiplicative model, the trend values are eliminated on dividing the given values (y) by the trend values (y_e). However, if we assume the additive model, the trend eliminated values are given by $(y - y_e)$. The resulting values contain short-term (cyclic) variations and irregular variations. Since the data are annual, the seasonal variations are absent.

ELIMINATION OF TREND

Year	*Trend eliminated values (in '000 tons) based on*	
	Additive Model $(y - y_e)$	*Multiplicative Model* (y / y_e)
1999	77 – 83 = – 6	77 ÷ 83 = 0·93
2000	88 – 85 = 3	88 ÷ 85 = 1·04
2001	94 – 87 = 7	94 ÷ 87 = 1·08
2002	85 – 89 = – 4	85 ÷ 89 = 0·96
2003	91 – 91 = 0	91 ÷ 91 = 1·00
2004	98 – 93 = 5	98 ÷ 93 = 1·05
2005	90 – 95 = –5	90 ÷ 95 = 0·95

Problem. *Fit a second degree parabola to the following data.*

X :	*1*	*2*	*3*	*4*	*5*
Y :	*1090*	*1220*	*1390*	*1625*	*1915*

Solution.

CALCULATIONS FOR PARABOLIC TREND

X	Y	$U = X - 3$	$V = \frac{Y}{1450}$	U^2	U^3	U^4	UV	U^2V
1	1090	–2	–72	4	– 8	16	144	–288
2	1220	–1	– 46	1	–1	1	46	–46
3	1390	0	–12	0	0	0	0	0
4	1625	1	35	1	1	1	35	35
5	1915	2	93	4	8	16	186	372
Total		$\Sigma U = 0$	$\Sigma V = -2$	$\Sigma U^2 = 10$	$\Sigma U^3 = 0$	$\Sigma U^4 = 34$	$\Sigma UV = 411$	$\Sigma U^2V = 73$

Let the parabola of best fit of V on U be :

$$V = a + bU + cU^2 \qquad \ldots(i)$$

where $U = X - 3$ and $V = \frac{Y - 1450}{5} \ldots(ii)$

Then the normal equations for estimating a, b and c are :

$$\left.\begin{aligned} \text{å}V &= na + b\text{å}U + c\text{å}U^2 \\ \text{å}UV &= a\text{å}U + b\text{å}U^2 + c\text{å}U^3 \\ \text{å}U^2V &= a\text{å}U^2 + b\text{å}U^3 + c\text{å}U^4 \end{aligned}\right\}$$

$$-2 = 5a + 10c \qquad \ldots(iii)$$

$$411 = 10b \qquad \ldots(iv)$$

$$73 = 10a + 34c \qquad \ldots(v)$$

(iv) $b = \frac{411}{10} = 41{\cdot}1 \qquad \ldots(vi)$

(v) $-2 \times (iii)$ gives : $73 + 4 = 34c - 20c = 14c \qquad c = \frac{77}{14} = 5{\cdot}5 \qquad \ldots(vii)$

Substituting the value of c in (iii), we get

$$5a = -2 - 10\,(5{\cdot}5) = -57 \qquad a = \frac{-57}{5} = -11\frac{4}{5} \qquad \ldots(viii)$$

Substituting the values of a, b, c from (vi), (vii) and $(viii)$ in (i), the parabola of best fit of V on U becomes :

$$V = -11{\cdot}4 + 41{\cdot}1\,U + 5{\cdot}5\,U^2 \qquad \ldots(ix)$$

where U and V are given by (ii).*

Substituting the values of U and V from (ii) in (ix), the second degree parabola of best fit of Y on X becomes

$$\frac{Y - 1450}{5}$$

$$= -11{\cdot}4 + 41{\cdot}1\,(X - 3) + 5{\cdot}5\,(X - 3)^2$$

$$= -11{\cdot}4 + 41{\cdot}1\,X - 123{\cdot}3 + 5{\cdot}5\,(X^2 - 6X + 9)$$

$$= 5{\cdot}5X^2 + (41{\cdot}1 - 33)\,X + (-11{\cdot}4 - 123{\cdot}3 + 49{\cdot}5)$$

$$= 5{\cdot}5X^2 + 8{\cdot}1X - 85{\cdot}2$$

$$Y - 1450 = 27{\cdot}5X^2 + 40{\cdot}5X - 426$$

$$Y = 27{\cdot}5X^2 + 40{\cdot}5X + 1024$$

Problem *The prices of a commodity during 2001—2006 are given below. Fit a parabola* $Y = a + bX + cX^2$ *to these data. Estimate the price for the year 2007 :*

Year (X) :	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Price (Rs.) (Y) :	100	107	128	140	181	192

Solution. Here, the number of pairs of observations $n = 6$ *i.e.,* even. Hence, shifting the origin to the arithmetic mean of two middle years, let us take :

$$t = \frac{X - \frac{1}{2}(2003 + 2004)}{\frac{1}{2}\,(\text{Interval})} = \frac{X - 2003{\cdot}5}{\frac{1}{2} \times 1} = 2(X - 2003{\cdot}5),$$

where X : Years ; Y : Price of commodity (in Rs.).

The values of t for $X = 2001$ to 2006 [From (*)] are respectively $-5, -3, -1, 1, 3, 5$.

Let the parabolic trend equation of Y on t be :

$$Y = a + bt + ct^2 ; \quad t = X - 2003{\cdot}5$$

where t unit $= \frac{1}{2}$ year and Y is price of the commodity in Rs.

CALCULATIONS FOR PARABOLIC TREND

Year (X)	*Price (in (Rs.) Y*	*t*	t^2	t^3	t^4	*ty*	t^2y
2001	100	–5	25	–125	625	–500	2500
2002	107	–3	9	–27	81	–321	963
2003	128	–1	1	–1	1	–128	128
2004	140	1	1	1	1	140	140
2005	181	3	9	27	81	543	1629
2006	192	5	25	125	625	960	4800
$n = 6$	$\Sigma Y = 848$	$\Sigma t = 0$	$\Sigma t^2 = 70$	$\Sigma t^3 = 0$	$\Sigma t^4 = 1414$	$\Sigma tY = 694$	$\Sigma t^2Y = 10160$

The normal equations for estimating *a, b* & *c* in (**) are :

$$\Sigma Y = na + b\Sigma t + c\Sigma t^2 \qquad 848 = 6a + 70c \qquad \ldots(i)$$

$$\Sigma tY = a\Sigma t + b\Sigma t^2 + c\Sigma t^3 \qquad 694 = 70b \qquad \ldots(ii)$$

$$\Sigma t^2Y = a\Sigma t^2 + b\Sigma t^3 + c\Sigma t^4 \qquad 10160 = 70a + 1414c \qquad \ldots(iii)$$

(*ii*) $$b = \frac{694}{70} = 9{\cdot}914 \qquad \ldots(iv)$$

Multiplying (*i*) by 35 and (*iii*) by 3 and then subtracting,

we get $29680 - 30480 = (210a + 2450c) - (210a + 4242c)$ $\quad -800 = -1792c$

$$c = \frac{800}{1792} = 0{\cdot}446 \qquad \ldots(v)$$

Substituting the value of *c* in (*i*), we get

$$848 = 6a + 31{\cdot}22 \text{ Þ } \quad a = \frac{848 - 31{\cdot}22}{6} = \frac{816{\cdot}78}{6} = 136{\cdot}130 \qquad \ldots(vi)$$

Substituting the values of *a, b* and *c* from (*iv*), (*v*) and (*vi*) in (**), we get the parabolic trend equation as :

$$Y = 136{\cdot}130 + 9{\cdot}914t + 0{\cdot}446t^2 \quad ; \quad t = 2(X - 2003{\cdot}5) \qquad \ldots(vii)$$

Estimation of price for 2007

When $X = 2007$, $t = 2(2007 - 2003{\cdot}5) = 2 \times 3{\cdot}5 = 7$.

Putting $t = 7$ in (*vii*), the estimated price of the commodity for the year 2007 is

$$\hat{Y}_{x=2007} = \hat{Y}_{t=7} = \text{Rs. } (136{\cdot}130 + 9{\cdot}914 \times 7 + 0{\cdot}446 \times 7^2)$$

$= \text{Rs. } (136{\cdot}130 + 69{\cdot}396 + 21{\cdot}854) = \text{Rs. } 227{\cdot}38$

Problem . *Fit a trend function* $y = A \cdot B^x$ *to the following data.*

x :	*1*	*2*	*3*	*4*	*5*
y :	*1·6*	*4·5*	*13·8*	*40·2*	*125·0*

Solution. Here we have to fit the exponential curve

$$y = A \cdot B^x \qquad \ldots(i)$$

Taking logarithm of both sides, we get

	$\log y$	$= \log A + x \log B$		
Þ	Y	$= a + bx$		… (*ii*)
where	$Y = \log y$;	$a = \log A$ and	$b = \log B$	… (*iii*)

Equation (*ii*) is straight line between the variables Y and x and hence the normal equations for estimating a and b are :

$$\text{å}Y = na + b\text{å}\,x \qquad \text{and} \qquad \text{å}\,xY = a\text{å}\,x + b\text{å}\,x^2 \qquad \ldots(iv)$$

x	y	$Y = \log y$	xY	x^2	*Trend Values* (y_e)	
1	1·6	0·2041	0·2041	1	1·5573	1·6
2	4·5	0·6532	1·3064	4	4·6361	4·6
3	13·8	1·1399	3·4197	9	13·8017	13·8
4	40·2	1·6042	6·4168	16	41·0877	41·1
5	125·0	2·0969	10·4845	25	122·3180	122·3
å $x = 15$		å$Y = 5·6983$	å $xY = 21·8315$	å $x^2 = 55$		

COMPUTATION OF EXPONENTIAL TREND

Substituting in (*iv*), the normal equations for estimating a and b become :

$$5·6983 = 5a + 15b \qquad \ldots(v) \qquad \text{and} \qquad 21·8315 = 15a + 55b \qquad \ldots(vi)$$

(iv) $-3 \times (v)$ gives :

$$21·8315 - 3 \times 5·6983 = 55b - 45b \text{ Þ } 10b = 4·7366 \text{ Þ } b = \frac{4·7366}{10} = 0·4737$$

Substituting the value of b in (*v*), we get

$$a = \frac{1}{5}(5·6983 - 15 \times 0·4737) = \frac{1}{5}(5·6983 - 7·1055) = -\frac{1}{5} \times 1·4072 = -0·2814$$

Hence using (*iii*), we get $B = \text{Antilog}\,(b) = \text{Antilog}\,(0·4737) = 2·977$

$A = \text{Antilog}\,(a) = \text{Antilog}\,(-0·2814) = \text{Antilog}\,(\bar{1}·7186) = 0·5231$

Substituting the values of A and B in (*i*), the required trend function is given by :

$$y = 0·5231 \times (2·977)^x \qquad \ldots(vii)$$

Putting x = 1, 2, 3, 4 and 5 in (*vii*), we get the trend values which are shown in the last column of the above table For example,

$(y_e)_1 = 0·5231 \times 2·977 = 1·5573$;

$(y_e)_2 = 0·5231 \times (2·977)^2 = 1·5573 \times 2·977 = 4·6361$

$(y_e)_3 = 4·6361 \times 2·977 = 13·8017$; and so on.

Problem *You are given the population figures of India as follows* :

Census Year (x) :	*1911*	*1921*	*1931*	*1941*	*1951*	*1961*	*1971*
Population (in crores) (y) :	*25·0*	*25·1*	*27·9*	*31·9*	*36·1*	*43·9*	*54·7*

Fit an exponential trend $y = ab^x$ *to the above data by the method of least squares and find the trend values. Estimate the population in 1981.*

Solution. Here $n = 7$, is odd. Further, since the population figures are given at equal intervals of 10 years, we define :

$$u = \frac{x - \text{middle value}}{\text{Interval}} = \frac{x - 1941}{10} \ldots (i)$$

and consider the trend curve ; $y = a \,.\, b^{u} \ldots (ii)$ Taking logarithm of both sides :

$$\log y = \log a + u \log b \Rightarrow Y = A + Bu \quad \ldots (iii)$$

where $Y = \log y$, $A = \log a$, $B = \log b$

The normal equations for estimating A and B in (iii) are given by (since $\sum u = 0$) : $\sum Y = nA$ and $\sum uY = B\sum u2 \Rightarrow A = \sum Y\, n$ and $B = \sum uY \sum u$

Year (x)	*Population* (*in crores*) (y)	$u = \frac{x - 1941}{10}$	$Y = \log y$	u^2	uY	*Trend Value* (*in crores*) $y_e = 33{\cdot}6 \times (1{\cdot}142)^{u}$
1911	25·0	–3	1·3979	9	– 4·1937	25·76 ÷ 1·142 = 2·56
1921	25·1	–2	1·3997	4	–2·7994	29·42 ÷ 1·142 = 25·76
1931	27·9	–1	1·4456	1	–1·4456	33·60 ÷ 1·142 = 29·42
1941	31·9	0	1·5038	0	0	33·60
1951	36·1	1	1·5575	1	1·5575	33·6 × 1·142 = 38·37
1961	43·9	2	1·6425	4	3·2850	38·37 × 1·142 = 43·82
1971	54·7	3	1·7380	9	5·2140	43·82 × 1·142 = 50·04
Total	$\sum y = 244{\cdot}6$	0	$\sum Y =$ 10·6850	$\sum u^2 =$ 28	$\sum uY =$ 1·6178	$\sum y_e = 243{\cdot}57$

Using (*iv*), we get $A = \frac{10{\cdot}6850}{} = 1{\cdot}5264$ Þ $a = \text{Antilog}\,(A) = \text{Antilog}\,(1{\cdot}5264) = 33{\cdot}60$

$B = \frac{1{\cdot}6178}{} = 0{\cdot}0578$ Þ $b = \text{Antilog}\,(B) = \text{Antilog}\,(0{\cdot}0578) = 1{\cdot}142$

Substituting the values of a and b in (*ii*), we get the exponential trend equation as :

$y = 33{\cdot}60 \times (1{\cdot}142)^{u}$, where $u = \frac{(x - 1941)}{10} \ldots (v)$

The trend values for the years 1911 to 1971 can be obtained from (*v*) on putting $u = -3, -2, \ldots, 2, 3$ respectively. For instance,

$$(y_e)_{x\,=\,1941} = (y_e)_{u\,=\,0} = 33{\cdot}60$$

Since the trend values given by (*v*) are in G.P. with common ratio $r = b = 1{\cdot}142$, the trend values for years 1951, 1961 and 1971 are obtained on multiplying 33·60 by 1·142 successively and similarly the trend values for 1931, 1921 and 1911 are obtained on dividing 33·60 by 1·142 successively. The trend values are given in the last column of Table 11·12.

Estimate of Population in 1981. For $x = 1981$, we get $u = \frac{x - 1941}{10} = \frac{1981 - 1941}{10} = 4$.

Hence putting $u = 4$ in (v), we get the estimated population of 1981 as : $(y_e)_{1981} = 33{\cdot}6 \times (1{\cdot}142)^4 = (33{\cdot}6) \times (1{\cdot}142)^3 \times 1{\cdot}142$

$= (y_e)_{1971} \times 1{\cdot}142 = 50{\cdot}04 \times 1{\cdot}142 = 57{\cdot}15$ (crores).

Or

$(y_e)_{1981} = 33{\cdot}6 \times (1{\cdot}30416)^2 = 33{\cdot}6 \times 1{\cdot}700844 = 57{\cdot}15$ (crores).

14.2.METHOD OF MOVING AVERAGES:

Method of moving averages is a very simple and flexible method of measuring trend. It consists in obtaining a series of moving averages, (arithmetic means), of successive overlapping groups or sections of the time series. The averaging process smoothens out fluctuations and the ups and downs in the given data. The moving average is characterised by a constant known as the *period* or *extent* of the moving average. Thus, the moving average of period '*m*' is a series of successive averages (A.M.'s) of *m* overlapping values at a time, starting with 1st, 2nd, 3rd value and so on.

Case (*i*) **When *Period is Odd.*** If the period '*m*' of the moving average is odd. then the successive values of the moving averages are placed against the middle values of the corresponding time intervals. For example, if *m* =5, the first moving average value is placed against the middle period. *i.e.,* 3rd, the second M.A. value is placed against the time period 4 and so on.

Case (*ii*). **When Period is Even.** If the period '*m*' of the M.A. is even, then there are two middle periods and the M.A. values are placed in between the two middle periods of the time intervals it covers. Obviously, in this case, the M.A. values will not coincide with a period of the given time series and an attempt is made to synchronise them with the original data by taking a two-period average of the moving averages and placing them in between the corresponding time periods. This technique is called *centering* and the corresponding moving average values are called *centred moving averages*. In particular, if the period *m* = 4, the first moving average value is placed against the middle of 2nd and 3rd time intervals; the second moving average value is placed in between 3rd and 4th time periods and so on.

If the time series data does not contain any movements except the trend which when plotted on a graph gives a straight line curve, then the moving average will reproduce the series. The following example will clarify the point.

Year	*Values*	*3-Yearly*	*5-Yearly*	*7-Yearly.*
(*t*)	(*y*)	*M.A.*	*M.A.*	*M.A.*
1	10	—	—	—
2	14	14	—	—
3	18	18	18	—
4	22	22	22	22
5	26	26	26	26
6	30	30	30	30
7	34	34	34	34
8	38	38	38	38

9	42	42	42	—
10	46	46	—	—
11	50	—	—	—

Thus the trend values by the moving average of extent 3, 5, 7 and so on coincide with the original series.

Note that in this case, the given values exhibit a linear trend $y = 6 + 4t$.

Moving Average and Curvilinear Trend.

If the data does not contain any oscillatory or irregular movements and has only general trend and the historigram (graph) of the time series gives a curve which is convex (concave) to the base, then the trend values computed by moving average method will give another curve parallel to the given curve but above (below) it. In other words, if there are no variations in the data except the trend which is curvilinear, then the moving average values, when plotted, will exhibit the same curvilinear pattern but slightly away from the given historigram. Further, *greater the period of the moving average, the farther will be trend curve from the original historigram.* In other words, the difference between the trend values and the original values becomes larger as the period of the moving average increases.

The moving average will completely eliminate the oscillatory movements if :

(i) The period of the moving average is equal to or a multiple of the period of oscillatory movements provided they are regular in period or amplitude, and

(ii) The trend is linear or approximately so.

Hence, to compute correct trend values by the method of moving averages, the *period or extent of the moving average should be same as the period of the cyclic movements in the series.* However, if the period of moving average is less or more than the period of the cyclic movement then it (M.A.) will only reduce their effect.

Quite often, we come across time series data which do not exhibit regular cyclic movements and might reflect different cycles with varying periods which may be determined on drawing the historigram of the given time series and observing the time distances between various peaks. In such a situation, the period of the moving average is taken as the average period of the various cycles present in the data.

Moving Average and Polynomial Trend. In most of the economic and business time series the trend is rarely linear and accordingly, if the trend is curvilinear, the moving average values will give a distorted picture of the trend. In such a case the correct trend values are obtained by taking a weighted moving average of the given values. The weights to be used will depend on the period of the M.A. and the degree of the polynomial trend to be fitted. For example, the weights for a moving average [5, 2], *i.e.,* a moving average of extent 5 for a parabolic trend are given by :

$$\left(-\frac{3}{35}, \frac{12}{35}, \frac{17}{35}, \frac{12}{35}, -\frac{3}{35}\right). \quad \ldots (*)$$

Thus, the first moving average value for series $y_1, y_2, y_3, \ldots$ is given by :

$$\frac{1}{35}(-3y_1 + 12y_2 + 17y_3 + 12y_4 - 3y_5).$$

The weights for the moving average [7, 2], *i.e.,* a M.A. of period 7 for parabolic trend are : $\left(-\frac{2}{21}, \frac{3}{21}, \frac{6}{21}, \frac{7}{21}, \frac{6}{21}, \frac{3}{21}, -\frac{2}{21}\right) \ldots (**)$ and the first trend value is given by :

$$\frac{1}{21}[-2y_1 + 3y_2 + 6y_3 + 7y_4 + 6y_5 + 3y_6 - 2y_7]$$

It may be observed that :

(iii) the weights for the M.A. are symmetric about the middle value, and

(iv) the sum of weights is unity.

Effect of Moving Average on Irregular Fluctuations. The moving average smoothens the ups and downs present in the original data and, therefore, reduces the intensity of irregular fluctuations to some extent. It can't eliminate them completely. However, greater the period of the moving average (up to a certain limit), the greater is the amount of reduction in their intensity. Thus, from point of view of reducing irregular variations, long-period moving average is recommended. However, we have pointed out in Remark 2, that greater the period of moving average, farther are the trend values from the original values. In other worlds, longer period of moving average is likely to give a distorted picture of the trend values. Accordingly, as a compromise, the period of moving average should neither be too large nor too small. *The optimum period of the moving average is the one that coincides with or is a multiple of the period of the cycle in the time series as it would completely eliminate cyclical variations, reduce the irregular variations and, therefore, give the best possible values of the trend.*

We shall now discuss numerical problems to explain the technique of obtaining trend values by moving average method.

Problem *Calculate (i) three yearly(ii) five yearly, moving averages for the following data and comment on the results.*

Year :	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
y :	*242*	*250*	*252*	*249*	*253*	*255*	*251*	*257*	*260*	*265*	*262*

Solution. The 3 yearly and 5 yearly moving average values are given in Table 11·14.

COMPUTATION OF 3 AND 5-YEARLY M.A. VALUES

Year *(1)*	*y* *(2)*	*3-yearly moving totals* *(3)*	*3-yearly moving averages (Trend values)* *(4) = (3) ÷ 3*	*5-yearly moving totals* *(5)*	*5-yearly M.A. (Trend values)* *(6) = (5) ÷ 5*
1990	242	—	—	—	—
1991	250	744	248·0	—	—
1992	252	751	250·3	1246	249·2
1993	249	754	251·3	1259	251·8
1994	253	757	252·3	1260	252·0
1995	255	759	253·0	1265	253·0
1996	251	763	254·3	1276	255·2
1997	257	768	256·0	1288	257·6
1998	260	782	260·7	1295	259·0
1999	265	787	262·3	—	—
2000	262	—	—	—	—

Comments. As the period of the M.A. increases, the trend values move away from the original values.

Problem. *Calculate the trend values by the method of moving average, assuming a four-yearly cycle, from the following data relating to sugar production in India+ :*

Year	*Sugar Production (lakh tonnes)*	*Year*	*Sugar Production (lakh tonnes)*
1971	*37·4*	*1977*	*48·4*
1972	*31·1*	*1978*	*64·6*
1973	*38·7*	*1979*	*58·4*
1974	*39·5*	*1980*	*38·6*
1975	*47·9*	*1981*	*51·4*
1976	*42·6*	*1982*	*84·4*

Solution. Since we are given that the data follows a four yearly cycle, we shall compute the trend values by using moving average of period 4, as shown in Table 11·15

COMPUTATION OF 4 - YEARLY MOVING AVERAGES

Year (1)	*Sugar production (lakh tonnes)* (2)	*4-yearly moving totals* (3)	*4-yearly moving average* (4) = (3) , 4	*2-period moving total of col. (4)* (5)	*Centred moving average [Trend values]* (6) = (5) , 2
1971	37·4				
1972	31·1				
1973	38·7	¬ 146·7	36·675	¬ 75·975	37·99
1974	39·5	¬ 157·2	39·300	¬ 81·475	40·74
1975	47·9	¬ 168·7	42·175	¬ 66·755	43·39
1976	42·6	¬ 178·4	44·600	¬ 95·475	47·74
1977	48·4	¬ 203·5	50·875	¬ 104·375	52·19
1978	64·6	¬ 214·0	53·500	¬ 106·000	53·00
1979	58·4	¬ 210·0	52·500	¬ 105·750	52·88
1980	38·6	¬ 213·0	53·250	¬ 111·450	55·73
		¬ 232·8	58·200		
1981	51·4				
1982	84·4				

Problem *Determine the period of the moving average for the following data and calculate moving averages for that period :*

Year	*:1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Value	*: 130*	*127*	*124*	*135*	*140*	*132*	*129*	*127*	*145*	*158*	*153*	*146*	*145*	*164*	*170*

COMPUTATION OF FIVE-YEARLY MOVING AVERAGE

Year *(1)*	*Value* *(2)*	*5-yearly Moving totals (3)*	*5-yearly M.A. (Trend Values) (4) = (3) , 5*
1	130	—	—
2	127	—	—
3	124	656	131·2
4	135	658	131·6
5	140	660	132·0
6	132	663	132·6
7	129	673	134·6
8	127	691	138·2
9	145	712	142·4
10	158	729	145·8
11	153	747	149·4
12	146	766	153·2
13	145	778	155·6
14	164	—	—
15	170	—	—

Solution. Since the peaks of the given data occur at the years 1, 5, 10 and 15, the data clearly exhibits a regular cyclic movement with period 5. Hence, the period of the moving average for determining the trend values is also 5, *viz.,* the period of the cyclic variations.

Merits and Demerits of Moving Average Method

Merits

1. This method does not require any mathematical complexities and is quite simple to understand and use as compared with the principle of least squares method.
2. Unlike the 'free hand curve' method, this method does not involve any element of subjectivity since the choice of the period of moving average is determined by the oscillatory movements in the data and not by the personal judgement of the investigator.
3. Unlike the method of trend fitting by principle of least squares, the moving average method is quite flexible in the sense that a few more observations may be added to the given data without affecting the trend values already obtained. The addition of some new observations will simply result in some more trend values at the end.
4. The oscillatory movements can be completely eliminated by choosing the period of the M.A. equal to or multiple of the period of cyclic movement in the given series. A proper choice of the period also reduces the irregular fluctuations to some extent.
5. In addition to the measurement of trend, the method of moving averages is also used for measurement of seasonal, cyclical and irregular fluctuations.

Limitations.

1. An obvious limitation of the moving average method is that we cannot obtain the trend values for all the given observations. We have to forego the trend values for some

observations at both the extremes (*i.e.,* in the beginning and at the end) depending on the period of the moving average. For example, for a moving average of period 5, 7 and 9, we lose the trend values for the first and last 2, 3 and 4 values respectively.

2. Since the trend values obtained by moving average method cannot be expressed by any functional relationship, this method cannot be used for forecasting or predicting future values which is the main objective of trend analysis.
3. The selection of the period of moving average is very important and is not easy to determine particularly when the time series does not exhibit cycles which are regular in period and amplitude. In such a case the moving average will not completely eliminate the oscillatory movements and consequently the moving average values will not represent a true picture of the general trend. [See Remark 3, § 11·5·6 for determining the period of M.A.]
4. In case of non-linear trend, which is generally the case in most of economic and business time series, the trend values given by the moving average method are biased and they lie either above or below the true sweep of the data. According to Waugh :

"If the trend line is concave downwards (like the side of a bowl), the value of the moving average will always be too high, if the trend is concave upward (like the side of a derby pot), the value of the moving average will always be tool low.

14.3 SUMMARY:

When we try to analyse the time series, we try to isolate and measure the effects of various kinds of these components on a series.
We have two models for analysing time series:

- Additive model, which considers the sum of various components resulting in the given values of the overall time series data and symbolically it would be expressed as: $Y = T + C + S + I$.
- The multiplicative model assumes that the various components interact in a multiplicative manner to produce the given values of the overall time series data and symbolically it would be expressed as :
  $y = T \times C \times S \times I$.
- The trend analysis brings out the effect of long-term causes. There are different methods of isolating trends, among these we have discussed only two methods which are usually used in research work, i.e. free hand and least square methods
- Long-term predictions can be made on the basis of trends, and only the least square method of trend computation offers this possibility.

14.4 SELF ASSESSMENT QUESTIONS:

1) The production (in thousand tons) in a sugar factory during 1994 to 2001 has been as follows:

Year	1994	1995	1996	1997	1998	1999	2000	2001
Produ-ction	35	38	49	41	56	58	76	75

(*Hint:* The point of origin must be taken between 1997 and 1998).

i) Find the trend values by applying the method of least square.
ii) What is the monthly increase in production?

2) Estimate the production of sugar for the year 2008.e following data relates to a survey of used car sales in a city for the period 1993-2001. Predict sales for 2006 by using the linear trend equation.

Years	1993	1994	1995	1996	1997	1998	1999	2000	2001
Sales	214	320	305	298	360	450	340	500	520

14.5. SUGGESTED READINGS:

- Mentgomery, D.C. and L.A. Johnson, 1996, '*Forecasting and Time Series Analysis*' McGraw Hill : New York.
- Chandan, J.S., 2001, *Statistics for Business and Economics*, Vikas Publishing House Pvt. Ltd., New Delhi.
- Gupta, S.P. and H.P. Gupta, 2001, *Business Statistics*, S. Chand, New Delhi.

Dr. Naga Nirmala Rani

LESSON- 15

DECISION THEORY

OBJECTIVE:

The purpose of studying this chapter is :

- Discuss the meaning and types of decision theory;
- Explain the decision-making under uncertainty;
- Describe the decision-making under risk situations; and
- Explain the decision tree analysis.

STRUCTURE:

15.1 INTRODUCTION TO DECISION THEORY:

We, as a person, make thousands of decisions daily. At times these decisions matter most and can influence our future in the long run. All such decisions like selection of vehicle, purchase of plot, rent of a farm, sharing/stocks investment etc. are critical decisions and everyone would like to make the right choice out of the available options. Decision theory is described as a set of methods that help the decision-maker choose the optimal action from several alternatives.

This chapter will focus on different decision rules available to decision maker under two decision making environments:

1. Decision-making under uncertainty
2. Decision-making under risk.

The decision making under uncertainty can be the problems in which a course of action can result in several possible outcomes, where the probabilities are not assigned, as a decision-

maker does not know which outcomes can or will occur. For example, a farmer living in a village might know that the outcome of the monsoon this year could be heavy rainfall, moderate rainfall, low rainfall, and so on, but he does not have the actual probability of its occurrence as he does not have past information with him to predict it from. In case of making a decision under risks, the decision-maker would have some additional info in the form of probabilities, the probability of heavy, moderate, and low rainfall is that it is derived from past knowledge and records.

15.2 MAKING DECISIONS IN UNCERTAIN CONDITIONS:

In an environment of uncertainty, a farmer will know that crop yield might be high, moderate or low but will not have the data to formulate probabilities of the states that will emerge. It is important to highlight that these yields depend on the type of crops that the farmer uses. Now he has to choose between the three crops (A, B and C), he knows that the possible yield (High, Moderate and Low) will have the following payoff (profits in 1000's Rs) as mentioned in the below Table 15.1.

Table 15.1 Payoff matrix (Profits in 1000's Rs)

Alternate Strategies		States of Nature		
		High	**Moderate**	**Low**
Crops	A	Rs. 60	Rs. 42	Rs. – 10
	B	Rs. 85	Rs. 60	Rs. – 20
	C	Rs. 50	Rs. 25	Rs. – 12.5

The above table is known as a payout table, we i.e. will attempt to assist the farmer in decision making by showcasing how the following decision rules can be applied:

1. The Maximax Criterion
2. The Maximin Criterion
3. The Minimax regret Criterion
4. Laplace's Criterion

15.2.1 Maximax Criterion

The second one is known as the maximax criterion because it is optimistic rule and assume that the best case will happen in the future. Using the maximax criterion we illustrate how to use the payoff in Table 15.2 just described.

Table 15.2 Payoff matrix

Rs. 60	Rs. 42	Rs. – 10
Rs. 85	Rs. 60	Rs. – 20
Rs. 50	Rs. 25	Rs. – 12.5

It is to be followed the next number of Steps,

Step 1: Choose the maximum payoff for each alternative/strategy. We can see the result in the below Table 15.3

Table 15.3 Payoff table

Crops	Payoff
A	Rs. 60
B	Rs. 85
C	Rs. 50

Step 2 Disclaimer: This site is designed for training only. Thus we have chosen the maximum of maximum strategies.

15.2.2 Maximin Criterion

Function: Maximin criterion: It is a pessimistic rule based on the assumption that the worst situation is likely to happen in the future and we would want to maximize the profits. We clarify the rule using the payoff matrix we just mentioned in Table 15.4:

Table 15.4 : Payoff matrix

Rs. 60	Rs. 42	Rs. – 10
Rs. 85	Rs. 60	Rs. – 20
Rs. 50	Rs. 25	Rs. – 12.5

Step 1: The minimum payoff for each strategy in following Table 15.5.

Table 15. 5 Payoff matrix

Crops	Payoff
A	Rs. – 10
B	Rs. – 20
C	Rs. – 12.5

Step 2: Choose the maximum gain of the minimum; crop A gives the maximum gain therefore going with crop A is advised

Note: It is also worthy to that in the event that we have a cost payoff matrix and not a payoff matrix for profit we will either be using the minimax criterion in which will be minimizing the maximum costs possible to incur. If given such a condition then the following steps are to be followed.

Step 1: Choose the highest cost of each strategy.

Step 2: Choose the lowest of the given maximum choice in step 1.

15.2.3 The Minimax Regrets Criterion

Minimax Regret Criterion: This method computes the opportunity loss or regrets for not taking the best decision for each state of nature, and the application of the minimax regret rule is presented below for our payoff matrix shown in Table 15.6

Table 15.6 : Payoff matrix

Rs. 60	Rs. 42	Rs. – 10
Rs. 85	Rs. 60	Rs. – 20
Rs. 50	Rs. 25	Rs. – 12.5

Step 1: Create a regret matrix, from the above payoff matrix, based on the following principle:

- When payoff represents profit
  i^{th} regret = (maximum payoff - i^{th} pay off) for the j^{th} event
- When payoff represents costs
  i^{th} regret = (maximum payoff - i^{th} pay off) for the j^{th} event

For the problem we choose for demonstration, the payoff matrix represents a profit payoff matrix. So the regret Table 15.7 are shown as below;

Table 15.7 Profit payoff matrix

Alternative	Pay off amounts				Regret Pay off amounts			Maximum Regret
	High	Medium	Low		High	Medium	Low	
A	Rs. 60	Rs. 42	Rs. - 10		25	18	0	25
B	Rs. 85	Rs. 60	Rs. - 20		0	0	0	10
C	Rs. 50	Rs. 25	Rs. – 12.5		35	35	2.5	35

From the regret payoff matrix, the minimum regret is for crops B (Regret = 10) which means the Crop B will be recommended.

15.2.4 Laplace Criterion

This is also called the equally likely decision criterion and it assumes that the states of nature (Low, Moderate, and High yield) will happen with equal likelihood. Then we calculate the expected value for each alternative. Below are the calculations pertaining to farmers' payoff matrix;

Crop A = 1/3 (60) + 1/3 (42) + 1/3 (−10) = 30.66

Crop B = (1/3)(85) + (1/3)(60) + (1/3)(−20) = 41.66

Crop C = 1/3 (50) + 1/3 (25) + 1/3 (−12.5) = 20.83

Crop B has the highest expected value and will thus be chosen by the farmer.

After showing the different rules available for decision-makers faced with a uncertain environments, the next step would be to analyze which of them is the best technique to apply in practice, since it is common to observe that the different techniques may lead to different results (i.e. selection of alternatives). Now, I would like to point out that the decision maker is free to adopt whether a particular technique is to his benefit or not based on his decision-making style, risk-taking ability, past behaviour, whether he knows the rule followed by his competitor or not, etc. As a consequence, different decision-makers are likely to utilize different rules for making decisions in a specific scenario.

15.3 DECISION MAKING UNDER RISK:

This type of risk environment assumes that the decision-maker has enough information about the different states of nature that applies to him such that he can assign the likelihood of occurrences for them as we noted in the introductory section of this unit. To follow along with our illustration, the farmer believes that with a probability of 0.65 a high yield will occur, 0.20 a moderate yield will occur, and with a probability of 0.15 a low yield will occur. From there, we can calculate the expected value for each alternative, as in Table 15.8.

Table 15.8 Payoff matrix

Alternate Strategies	States of Nature			Expected Value
	High yield	Moderate yield	Low yield	
Probabilities	*0.65*	*0.20*	*0.15*	
Crop A	Rs. 60	Rs. 42	Rs. – 10	45.90
Crop B	Rs. 85	Rs. 60	Rs. – 20	64.25
Crop C	Rs. 50	Rs. 25	Rs. – 12.5	35.62

Expected value for
Crop A = 0.65 ´ Rs. 60 + 0.20 ´ Rs. 42 + 0.15 ´ Rs. (10) = 45.90

Expected value for
Crop B = 0.65 ´ Rs. 85 + 0.20 ´ Rs. 60 + 0.15 ´ Rs. (20) = 64.25

Expected value for
Crop C = 0.65 ´ Rs. 50 + 0.20 ´ Rs. 25 + 0.15 ´ Rs. (12.5) = 35.62

In this case, the farmer would now choose crop B to ensure a greater expected value.

Let us say this is a common scenario because in daily lives we make decisions. Naresh is a newspaper seller who delivers newspapers in a small society. He purchases the newspaper from a local vendor at Rs. 1 unit and sells it at Rs. 1.2 each paper. Any paper that remains unsold by the end of the day is valued at Rs. 0.30 per kg and sold off to a scrap dealer. Its demand varied from 60 to 65 customers a day. He also now loses customers if he does not

have enough stock. Naresh has kept a record of his customers. His data shows that the demand for this newspaper over the last 300 days was as highlighted in Table 15.9

Table 15.9 300 days demand table.

Demand in units	60	61	62	63	64	65
No. of days	45	60	90	60	30	15
Probability of demand (i.e. no. of days / 300)	0.15	0.20	0.30	0.20	0.10	0.05

To find the probability that the demand is 60, 61, 62 [...] and so on, we can compute the relative frequency here. Payoff Table 15.10 (below) is constructed now.

Table 15.10 Payoff matrix

Stocking Policy	**60**	**61**	**62**	**63**	**64**	**65**
Probability of demand	0.15	0.20	0.30	0.20	0.10	0.05
60	12	12	12	12	12	12
61	11.1	12.2	12.2	12.2	12.2	12.2
62	10.2	11.3	12.4	12.4	12.4	12.4
63	9.3	10.4	11.5	12.6	12.6	12.6
64	8.4	9.5	10.6	11.7	12.8	12.8
65	7.5	8.6	9.7	10.8	11.9	13.0

As shown in the above payoff table, Naresh's profit keeps changing depending on his stocking units for that day and the demand which varies between 60 to 65. As observed If Naresh inventories 63 units and the demand on that day is 63 units, the earnings in profit will be;

63 units x 0.20 (profit margin) = 12.6

However, if on that day demand is below his stocking level of 63 units and he sells only 60 units then his profit will be;

(profit from the sold unit) - loss from unsold units (i.e. Rs. 0.90 per unit)
= (60 units x 0.2 profit/unit) - (3 units x 0.90 / unit)
= Rs. 12 - Rs. 2.7
= Rs. 9.3

For a demand of 65, if Naresh stocks 64 units he can earn maximum profit of Rs.12. 8 as all the units are sold. This is also called conditional table, which is table 15.10.

In the case of this problem, the conditional table 15.11 informs how much profits Naresh will be making under different stocking and demand conditions, but it does not inform you about how much is to be stocked by Naresh. Moving ahead we would like to utilize the given

probabilities of demand occurrence and compute expected profits for various stocking scenarios shown in table 9

Table 15.11: Expected Profit

	Demand						
Stocking Policy Probability	**60**	**61**	**62**	**63**	**64**	**65**	**Expected Value**
	0.15	**0.20**	**0.30**	**0.20**	**0.10**	**0.05**	
60	12	12	12	12	12	12	**12**
61	11.1	12.2	12.2	12.2	12.2	12.2	**12.035**
62	10.2	11.3	12.4	12.4	12.4	12.4	**11.3**
63	9.3	10.4	11.5	12.6	12.6	12.6	**11.335**
64	8.4	9.5	10.6	11.7	12.8	12.8	**10.6**
65	7.5	8.6	9.7	10.8	11.9	13.0	**9.755**

Expected values for;

60 units stocking policy
= 12 x 0.15 + 12 x 0.2 + 12x 0.3 + 12 x 0.2 + 12 x 0.1 + 12 x 0.05 = 12

61 unit stocking policy
= 11.1 x 0.15 + 12.2 x 0.2 + 12.2 x 0.3 + 12.2 x 0.2 + 12.2 x 0.1 + 12.2 x 0.05
= 12.035 and so on

62 unit stocking policy = 11.3

63 unit stocking policy = 11.335

64 unit stocking policy = 10.6

65 unit stocking policy = 9.7555

Naresh should stock 61 newspapers per day based on above situation expected values as this stock level provides him the maximum expected value. The following concept that we will be covering now is the Expected Value of Perfect Information (EVPI).

Where the decision-maker pays money to obtain information, so he has perfect knowledge of what will be the demand for that day, so he can stock only the required quantity, the definition of this concept is called the 'EVPI' as the name refers. Naresh can solve his problem if he gets information on the probability of occurrence of demand such that he will not lose Rs. 0.9 on every newspaper left unsold. The demand will still fluctuate between 60 -65 units per day for Naresh even if he has perfect information. The conditional profit table of Naresh with perfect information is given in Table 15.12. It indicates the maximum profit that he can receive.

Table 15.12 Conditional profit table

Stocking Policy Probability	Demand						Expected Value
	60	61	62	63	64	65	
	0.15	0.20	0.30	0.20	0.10	0.05	
60	12						**1.8**
61		12.2					**2.44**
62			12.4				**3.72**
63				12.6			**2.52**
64					12.8		**1.28**
65						13.0	**0.65**
						Total	**12.41**

The profit that Naresh can make with perfect information is maximum Rs.12. 41. So this information was not available, therefore Naresh's profit without this information would have been Rs. 12.035. The value of perfect information = Rs. 12.41 – Rs. 12.035 = Rs. 0.375 This suggests that Naresh should therefore not spend above Rs. 0.375 per day on gathering a perfect information otherwise he will not be able to make the maximum profit of Rs. 12.41.

15.4 DECISION MAKING UNDER CERTAINTY:

Decision Making under Certainty: The Decision maker has complete and precise information about all the possible alternatives and their results i.e Decision theory where the decision maker knows the precise results of each option. There is no uncertainty or risk—each action produces a concrete and predictable outcome.

The results of each option are entirely known.

Nothing is probabilistic or uncertain.

The decision-maker then chooses the option resulting in the most grossly perceived reward relevant to their values or objectives.

Example 1: Buying a Laptop

This is what happens folks: Imagine you store statistics until October 2023. You have three options:

Option	Laptop A	Laptop B	Laptop C
Price	$900	$1,100	$1,300
Battery Life	6 hrs	10 hrs	8 hrs

You just care about battery life, and you already know them for sure.

Decision: Select Laptop B as it has more battery life.

Example 2: How to Choose a Mode of Transportation

You're figuring how you get to work. Your options:

Option	Time Taken	Cost
Drive Yourself	30 minutes	$5 gas
Take the Bus	45 minutes	$2 fare
Ride a Bike	25 minutes	Free

If your objective is to minimize time and you know exactly how long each option takes,

Decide: go ride a bike (fastest and cheapest)

15.5 DECISION TREE ANALYSIS:

The above mentioned decision environments restrict the reference of decision making to whether the probabilities are available or not. In addition, neither of the environments consider the time series of the actions and resulting events would take place.

As we know the decision making involves multiple stages and at each stage, either of the alternative will lead to different events and payoffs, we present a tree formation listing all the possible events and resultant payoffs in a decision tree analysis.

Figure 15.1 shows different components of a decision tree

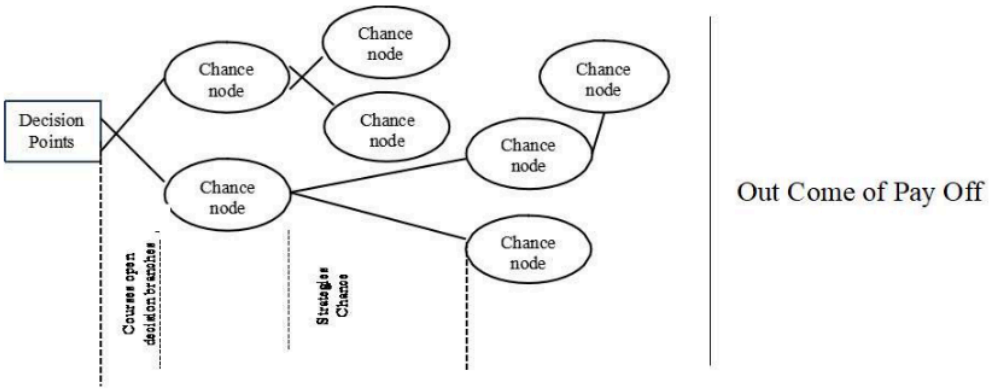


Fig. 15.1: Decision Tree

Decisions Points: The Square as a decision point The branches extending from this square show the different paths available to the decision-maker and are referred to as decision branches.

Node: Represented by a circle. Each branch ends in a node. Each node has different outcomes with the associated probabilities estimates comes out of it. Chance branches: Branches also emerge out of the nodes and are called as chance branches. The probability of which actions would be assumed in that solution are reflected in these chance branches. The payoff indicated in the previous figure taken against each chance branch can be positive or negative according

to the nature of the event. Positive payoffs represent profits or revenues and negative payoffs represent expenditure and losses.

Now, let us solve a decision problem with a decision tree using below illustrative example

Example 1

A farm equipment manufacturing firm tries to choose between two alternatives as below:

Alternative I: Revenue generation via dealers

Alternative II: The equipment is sold directly (own showroom)

Payoff when sale occurs when sold through dealer = Rs.70 lakhs Probability that sales are high giving payoff Rs.70 lakhs = 0.7 When the sales are low, the possible payoff is Rs. 30 lakhs. If - the company chooses the option of its own distribution (direct-selling) the probability of one having high sales is 0.60 with pay-off 90 lakhs and 25 lakhs respectively.

Now, let us represent the decision tree to gain an insight into the problem. We understand that it is a decision over two alternatives and therefore this can be represented as a decision point in following Figure 15.2

Fig.15.2 Decision Tree

Then, we have two outcomes for each alternative: high sales or low sales, with some expected payoff. And is depicted as follows in figure 15.3

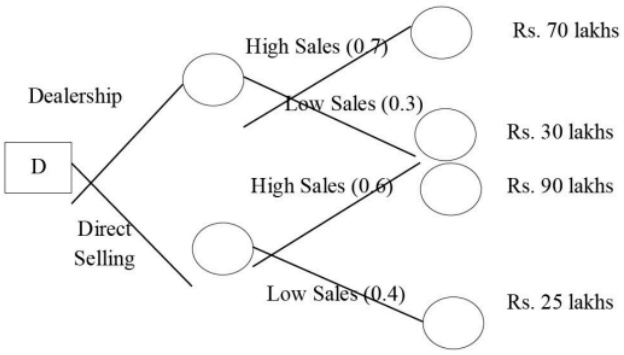


Fig.15.3 Decision Tree

15.6 SUMMARY:

Decision theory has emerged as a critical tool for making decisions under uncertainty. This unit covered in detail the decision theory approaches constructed to resolve problems with a small number of decision alternatives and a limited list of possible states of nature. The decision theory approach aimed to determine the best decision/operating alternatives in the presence of an uncertain or risk-laden future-state pattern (or states of nature).

Decision making under uncertainty I mentioned various decision techniques such as maximin, maximax, minimax regret and the Laplace criterion. The method applied is called decision-making under risk, and we find best decision and calculate expected value of perfect information when the problems have probabilities known to the decision-maker based on the previous dataset. The EVPI is the value of the perfect information that the decision maker can pay to know about the occurrence of an event.

Sequential decision-making scenarios can be analysed through decision trees. A graphical tree diagram organizes the decision alternatives and potential events. In this case, the tree is examined from right to left. At each event node the expected value computations are made, and at each decision points the most favorable decision alternatives are recognized.

15.7 TECHNICAL TERMS:

- **Decision Theory :** It is a method or framework of logical and mathematical concepts, aimed at helping decision makers to formulate rules and choose among a set of alternatives.
- **Decision Tree Analysis :** A schematic method of alternatives available to the decision-maker, to analysis the circumstances along with their possible consequences."
- **Laplace Criterion :** This method explicitly uses the probability of assessments regarding the likelihood of occurrence of the states of nature.
- **Maximax Criterion :** This method looks at the best that could happen under each action and then chooses the action with the largest value.
- **Maximin Criterion :** This method involves selecting the alternative that maximises the minimum pay-off achievable.
- **Minimax Regrets Criterion :** This method minimizes regret which is highest when one decision has been made instead of another.

15.8 SELF ASSESSMENT PROBLEMS:

1. What do you understand by the term decision theory?
2. Describe some methods which are useful for decision-making under uncertainty.
3. What techniques are used to solve decision-making problems under uncertainty?
4. Which decision theory techniques result in an optimistic decision? Which technique results in a pessimistic decision?
5. What do you understand by decision-making under risk?
6. Write a note on the value of perfect information.

7. What are the main steps associated with decision tree analysis.
8. Mr. Naresh wants to invest Rs. 10,000 in one of the three options A, B, and C. The payoff for his investment depends on the nature of the benefits he gets from the investment (share market, gold market, fixed deposit). The possible returns under each situation are given below;

Strategy	Nature of Benefits from the investment		
	Share Market(S_1)	Gold Market(S2)	Fixed Deposit(S_3)
A	2000	1200	1500
B	3000	800	1000
C	2500	1000	1800

What course of action has he to take according to,

i. Maximin
ii. Maximax
iii. Laplace
iv. Minimax (the regret criterion)

9. At a cake shop, special cookies are sold, it has the following probability of selling cookies.

No. of cookies sold	**Probability**
10	0.10
11	0.15
12	0.20
13	0.25
14	0.30

The cost of a cookie is 30 paise and the sale price is 50 paise. At end of the day, unsold cookies have to be thrown. How many cookies should the shop owner order to avoid wastage?

15.9 REFERENCES:

1. Sharma, A. (2009). Operations Research. Global Media, Himalaya Publishing House.
2. Sharma, J.K. (2010), Operations Research – Problems and Solutions, Third Edition, Macmillan Publishers India Ltd.
3. Taha, H. A. (2008), Operations Research – An Introduction, Eight Edition, Prentice – Hall of India Private Ltd.

Dr. P.Vijaya Vani

CHAPTER- 16

LINEAR PROGRAMMING PROBLEM

OBJECTIVE:

The purpose of studying this chapter is :

- To understand what is Linear Programming.
- To formulate Linear Programming Models.
- To understand Graphical Method of Solution.
- To solve two variable LP Problems by Graphical Method.

STRUCTURE:

16.1 LINEAR PROGRAMMING INTRODUCTION:

Linear Programming is a decision-making technique used to assist business executives. Linear Programming is a mathematical method for how to make the best use or to allocate the resources and to achieve the desired objective when there are several possible uses of the same resource, be it money, manpower, material, machine and other facilities.

16.1.1 The Nature of Linear Programming Problem

Two of the most common are:

1. The product-mix problem
2. The blending Problem

Product Mix Problem: In the product mix problem, it is the problem which involves two or more products also called candidates or activities which compete for limited resources. The challenge is determining which products and how much of them should be manufactured or sold as part of production schedule, in order to achieve maximum profit, market share or sales revenue. The blending problem is to find the optimum blend of available ingredients to produce a certain amount of product, within strict specifications (Riyad & Bilic, 2003). The best blend is the blend of required inputs at minimum cost.

16.1.2 Terminology Used in A Linear Programming Problem

1. **Elements of LP Problem:** Each LPP consists of a. Decision Variable, b. Objective Function, c. Constraints.
2. **Optimization:** Linear Programming, one can understand using values of linear equations under the constraints placed, we try to optimise either the maximum or minimum values of the objective function.
3. **Cost Coefficient of Profit:** The coefficient of the variable in objective function gives the expected value of increase or decrease of the objective function per unit increase in the solution.
4. **Constraints:** Maximising (or minimising) is done under a system of constraints. So LP can also be defined as a constrained optimisation problem. They are the limitations of the resources.
5. **Input-Output coefficients:** The coefficient of constraint variables are termed as Input- Output Coefficients. They show the pace at which a resource is consumed or utilised. They are found to the left-hand side of the constraints.
6. **Capacities:** They are in the form of inequalities and represent limits on the right hand side (RHS) of the resource use constraints.

16.1.3 The Mathematical Expression of The LP Model

The general LP Model can be mathematically expressed as: Let

O_{ij}= Input-Output Coefficient

c_j = Cost (Profit) Coefficient

b_i = Capacities (Right hand Side)

x_j = Decision Variables

Find a vector (x_1, x_2, x_3 .....x_n) that minimise or maximise a linear objective function Z(x)

where $Z(x) = c_1x_1 + c_2x_2 + \ldots\ldots\ldots\ldots + c_nx_n$

subject to linear constraints

$a_1x_1 + a_2x_2 + \ldots\ldots\ldots\ldots + a_nx_n = b_2$

$a_1x_1 + a_2x_2 + \ldots\ldots\ldots\ldots + a_nx_n \leq b_2$

.....

.....

.....

.....

$a_{m1}x_1 + a_{m2}x_2 + \ldots\ldots\ldots\ldots + a_{mn}x_n \leq b_m$ and

non-negativity constraints

$x_1 \geq 0, x_2 \geq 0, , x_n \geq 0$

16.1.4 Formulation of LPP Steps

1. Identify decision variables
2. Write objective function
3. Formulate constraints

Example 1. (Production Allocation Problem)

A company manufactures three products. Here I am just processing these products across three different machines. The following table presents the time taken to produce a single unit of each of the three products and the daily capacity of the three machines:

Machine	Time per unit (Minutes)			Machine Capacity (minutes/day)
	Product 1	Product 2	Product 3	
M1	2	3	2	440
M2	4	-	3	470
M3	2	5	-	430

Assuming that it is necessary to establish for each product the daily number of units that should be produced. Therefore the profit/unit is Rs. 4, Rs.3 and Rs. 6 for product 1, 2 and 3 respectively. It is assumed that all amounts are consumed in the market. Based on the problem statement, the mathematical (L.P.) model that will maximise the daily profit is formulated.

16.1.5 Formulation of Linear Programming Model

Step 1

Take the lessons from the study of the situation. In the situation, key decision is how much of the products 1, 2 and 3 to be produced, as the extents are allowed to differ.

Step 2

Notice variable quantities of interest in step 1, and assume symbols for them. Let x_1, x_2 and x_3 units be the extents (amounts) of products 1, 2 and 3 manufactured each day.

Step 3

Mathematically express the possible options in terms of variable. Feasible alternatives are physically, economically and financially possible. Feasible alternative in the given situation are values x_1, x_2 and x_3 units of x_1, x_2 and x_3 respectively where x_1, x_2 and $x_3 \geq 0$. Since producing negative units is infeasible and nonsensical.

Step 4

Disclose the goal function and express it quantitatively as a linear function of variables. Since in current scenario, the goal is to derive the maximum profit. i.e., $Z = 4x_1 + 3x_2 + 6x_3$

Step 5

So state the influencers or restrictions. These happen mainly due to limits on availability (resources) or needs (demands). Write these constraints as these linear equations/inequalities in variables.

Here, we know constraints should be on machine capacities which can be written mathematically as

$$2x_1 + 3x_2 + 2x_3 \leq 440$$
$$4x_1 + 0x_2 + 3x_3 \leq 470$$
$$2x_1 + 5x_2 + 0x_3 \leq 430$$

Example 2: Product Mix Problem

Use the following information to answer questions 1 to 3: A factory produces two types of goods A and B. To produce one unit of A, it takes 1.5 hours of machine time and 2.5 hours of labour time. In order to produce product B, it takes 2.5 hours of machine time and 1.5 hours

of labour time. 300 machine hours and 240 labour hours are available for a month. Total editing per unit profit for A is Rs. 50 and for B is Rs. 40. Formulate as LPP.

Solution:

Products	Resource/unit	
	Machine	Labour
A	1.5	2.5
B	2.5	1.5
Availability	300 hrs	240 hrs

There are going to be two constraints. Two for the availability of machine hours and the availability of labour hours.

Decision variables

x_1 = Units of A produced monthly.

x_2 = Units of B produced per month

The objective function: Max $Z = 50x_1 + 40x_2$

Subjective Constraints

For machine hours $1.5x_1 + 2.5x_2 \leq 300$

For labour hours $2.5x_1 + 1.5x_2 \leq 240$

Non negativity $x_1, x_2 \geq 0$

Example: 3

A company manufactures three products A, B, C. Three raw names used in the process of manufacturing P, Q and R. Unit Profit A - Rs. 5 B - Rs. 3 C - Rs. 4

Resource requirements/unit

Maximum available raw materials:

P – 80 units; Q – 100 units; R – 150 units. Formulate LPP.

Solution:

Raw Material / Product	P	Q	R
A	-	20	50
B	20	30	-
C	30	20	40

Decision variables:

x_1 = Number of units of A

x_2 = Number of units of B

x_3 = Number of units of C

Objective Function

As Profit per unit is provided, the objective function is maximisation

Max $Z = 5x_1 + 3x_2 + 4x_3$

Constraints:

For P: For Q: For R:

$0x_1 + 20x_2 + 30x_3 \leq 80$

$20x_1 + 30x_2 + 20x_3 \leq 100$

$50x_1 + 0x_2 + 40x_3 \leq 150$

(for B, R is not required)

$x_1, x_2, x_3 \geq 0$

Example 4: Portfolio Selection Investment Decision

We have an investor who wants to invest in two securities 'A' and 'B'. The risk-return profile of these securities are different. Security 'A' provides a return of 9% and a risk factor of 5 on a scale of zero to 10. Security 'B' behaves 15% return with 8 risk factor. We will invest a total amount of Rs. 5,00,000/- and need a minimum return of 12% on the investment. Maximum combined risk should not be more than 6. Formulate as LPP.

Solution:

Decision Variables:

x_1 = Amount invested in Security A

x_2 = Amount invested in Security B

Objective Function:

The aim is to optimize the return on the overall investment.

$\therefore$ Max Z = 0.09 x1 + 0.015 x2 (where % = 0.09 and 15% = 0.15)

Constraints:

1. Related to Total Investment:
   $x_1 + x_2 = 5,00,000$
2. Related to Risk:
   $5x_1 + 8x_2 = (6 \times 5,00,000)$
   $5x_1 + 8x_2 = 30,00,000$
3. Related to Returns:
   $0.09x_1 + 0.15x_2 = (0.12 \times 5,00,000)$
   $\therefore 0.09x_1 + 0.15x_2 = 60,000$
4. Non-negativity $x_1, x_2 \geq 0$

Example 5: Inspection Problem

Analysis of sample data A company performs quality control inspection with two grades of inspectors: I and II. At least 1, 500 pieces are to be inspected in one 8-hours long day. Inspector of grade I can inspect 20 pieces in one hour with 96% accuracy. Grade II inspector inspects 14 items per hour with 92% accuracy. The hourly wages of grade I inspector is not less than Rs. 5 while grade II inspector is not less than Rs. 4 A single error from an inspector costs the company Rs. 3. It is given that there are, in total, 10 grade I inspectors, and 15 grade II inspectors in the company. Determine the most efficient assignment of inspectors such that the daily inspection cost is minimized.

Solution:

Consider x1 and x2 be the number of grade I and grade II inspectors that can be assigned the job of quality control inspection.

The goal is to minimize the daily inspection cost. The company now incurs two types of costs; the cost of paying the inspectors and the cost of inspectors making a mistake.

Inspector/hour cost of grade I is

Rs. (5 + 3 X 0.04 X 20) = Rs.

Cost of grade II inspector/hour is likewise

= Rs. (4 + 3 X 0.08 X 14) = Rs. 7.36.

$\therefore$ The objective function is

Min $Z = 8 (7.40x_1 + 7.36x_2) = 59.20 x_1 + 58.88 x_2$.

We can write the constraints on the number of grade I inspectors as: $x_1 \leq 10$,

on the number of grade II inspectors: $x_2 \leq 15$

on how many to inspect every day:

$20 \times 8x_1 + 14 \times 8x_2 \geq 1500$

or $160x_1 + 112x_2 \geq 1500$

where, $x_1, x_2 \geq 0$.

Example 6: Trim Loss Problem

A cylindrical container manufacturer receives sheets of tin of sizes 30 cm and 60 cm in width. These containers have 3 different sizes.By dividing sheets into 3 sizes of 15 cm,21 cm & 27 cm, respectively. From these three widths the amount of containers to be manufactured are respectively 400, 200 and 300. In the original labourers, containers are bought from the market, its bottom plates and top covers. Standard tin sheets have no length limits. Minimising trim losses production schedule is your asked output (LPP formulation)

Solution:

That is, given the two standard sizes of tin sheets, a key decision is to decide how each of the standard sizes of tin sheets be cut to the require sizes such that the trim losses are minimum. There are combinations of the three needed widths of 15 cm, 21 cm and 27 cm from widths of 30 cm and 60 cm available.

We denote these combinations with x_{ij}. Trim loss occurs for each combination. Constraints can be formulated in the following way:

Below is the table showing the potential cutting combinations (plans) of both types of sheets:

Width(cm)	i = I (30 cm)			i = II (60 cm)					
	x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}
15	2	0	0	4	2	2	1	0	0
21	0	1	0	0	1	0	2	1	0
27	0	0	1	0	0	1	0	1	2
Trim Loss (cm)	0	9	3	0	9	3	3	12	6

Thus, the constraints are

$2x_{11} + 4x_{21} + 2x_{22} + 2x_{23} + x_{24} \geq 400$

$x_{12} + x_{22} + 2x_{24} + x_{25} \geq 200$

$x_{13} + x_{23} + x_{25} + x_{26} \geq 300$

Objective is to maximise the trim losses.

i. e., minimise $Z = 9x_{12} + 3x_{13} + 9x_{22} + 3x_{23} + 3x_{24} + 12x_{25} + 6x_{26}$

ii. where $x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{26} \geq 0$.

Example 7: Diet Problem

Two foods F1 and F2 contain vitamins B1 and B2 respectively. F1 uses 3 B1 and 4 B2 per unit. Now 1 unit of F2 comprises F1 + 3 B2. Minimum daily prescribed intake of B1 & B2 is 50 and 60 units respectively. F1 & F2 represent units with the cost per unit as Rs. 6 & Rs. 3 respectively. Formulate as LPP.

Solution:

Vitamins	Foods		Minimum Consumption
	F_1	F_2	
B_1	3	5	30
B_2	5	7	40

Decision Variables:

x1 = No. of units of P_1 per day.

x2 = No. of units of P_2 per day.

Objective function:

Min. $Z = 100\ x_1 + 150\ x_2$
Subject to constraints:
$3x_1 + 5x_2 \geq 30$ (for N_1)
$5x_1 + 7x_2 \geq 40$ (for N_2) $x_1, x_2 \geq 0$

Example 8: Blending Problem
An oil company manager is due to find an optimal blend of two blending pages. Formulate LPP.
Data:

Process	Input (Crude Oil)		Output (Gasoline)	
	Grade A	Grade B	X	Y
P1	6	4	6	9
P2	5	6	5	5

Profit per operation:
Process 1 (P1) = Rs. 4,000 Process 2 (P2) = Rs. 5,000
Crude oil maximum availability:
Grade A = 500 units Grade B = 400 units
Minimum Demand for Gasoline:
X = 300 units Y = 200 units
Solution:
Decision Variables:
x_1 = No. of operations of P1 x_2 = No. of operations of P2
Objective Function:
Max. $Z = 4000\ x_1 + 5000\ x_2$
Subjective to constraints:
$6x1 + 5x2 \leq 500$
$4x1 + 6x2 \leq 400$
$6x1 + 5x2 \geq 300$
$9x1 + 5x2 \geq 200$
$x1, x2 \geq 0$

16.1.6 Merits of LPP

1. Facilitates management in utilizing reusable resources effectively.
2. Provides quality for decision making.
3. Great instruments to adapt to the evolving requirements.
4. To use a computer to quickly determine the solution.
5. Yields a natural sensitivity analysis.
6. Finds solution to problems with a very large or infinite number of possible solution.

16.1.7 Demerits of LPP

1. **Non-linear equation exists:** The basic requirement of Linear Programming is that the objective function and constraint function must be linear. A few conditions applying to practically linear relationship do not hold true.
2. **Joint and non-linearity:** LP cannot handle joint behaviour among different variables, which result in a non-linear interaction in the equation between joint interaction between some of the activities like total effectiveness.
3. **Use of Fractional Value:** In LPP fractional values are allowed in decision variable.

4. **Information on Coefficients of the equation:** It is not always possible to claim all coefficients in the objective function and limitations with certainty.

16.1.8 Self Assessment Exercises

1. Define the following terms in the context of Linear Programming: decision variables, objective function and constraints.
2. Formulate the linear programming problems mathematically.
3. What does LPP consist of? Why the constraint of non-negativity is important?
4. State the limitations of LPP.
5. List the assumptions and benefits of LPP.
6. For example, an investor needs to determine what amount to be invested in one equity and one debt fund. Total Fund = Rs. 5,00,000. Invest in one fund no more than Rs. 3,00,000. Equity returns expected is 30% and debt 8%. Return on total investment, minimum is 15% Formulate as LPP.
7. A company produces two types of products P1 and P2. Profit per unit for P1 is Rs. 200 and for P2 is Rs. 300. M1, M2 and M3 are the three raw materials required. M1: 5 units, M2: 10 units (for 1 unit of P1) Unit 1: P2 requires 18, M2, 10, M3 Available are 50 units of M1, 90 units of M2 and 50 units of M3. Formulate as LPP.
8. A company manufactures two products X and Y. Minimum production of X must be 50 units. Y can be produced without limit. Profit per unit of X and Y is Rs. 100 and Rs. 150 respectively. Formulate as LPP.

Product	Resource Requirement	Resource Availability
X	20 Machine Hours	Machine Hours = 2500
	10 Labour Hours	Labour Hours = 3000
Y	10 Machine Hours	
	15 Labour Hours	

9. Two Nutrients N1 and N2 have been advised to a patient as daily consumption to consider. For N1, 10g; N2 needs a minimum intake of 15g per day Two products P1 and P2 are available in these nutrients. For example, 1 unit of P1 has 2g of N1 and 3g of N2. P2 is made up of 1g of N1 and 2g of N2. P1 and P2 have a cost per unit of Rs. 200 and Rs. 150 respectively. The nutrient requirement must be satisfied at minimum cost, formulate as LPP.
10. Vitamin A and B are going to be given daily to students as health supplements. Suppose you have two products Alpha & Beta that contain vitamins A and B. One unit of Alpha contains 2g of A and 1g of B. One unit of Beta contains 1g of A and 2g of B. Your daily requirement of A and B is at least 10g each. Cost per unit of Alpha is Rs.20 and of Beta is Rs.30. Maintain all requirements at the least possible cost and formulate it as LPP.

16.2 FOR LINEAR PROGRAMMING SOLUTION - GRAPHICAL METHOD:

16.2.1 Introduction

There are two approaches to finding an optimal solution in a Linear Programming Problem. The first is graphical method and second is simplex method.

Constructing a graphical method can be made only for a problem related to two variables i.e. there are two decision variables. The two decision variables x1 & x2 are plotted on the two axes of the graph (X & Y axis)

16.2.2 Maximisation Model

Step 1 : LPP formulation (Linear Programming Problem)
Based on the information provided, write the Linear Programming Problem (LPP).

Example

A firm produces two products A and B. Both the products are processed on two machines M1 & M2.

	M1	M2
A	6 Hrs/Unit	2 Hrs/Unit
B	4 Hrs/Unit	4 Hrs/Unit
Availability	7200 Hrs/month	4000 Hrs/month

For A, profit per unit = Rs. 100 and profit per unit for B is Rs. 80. Determine the monthly output of A and B that will maximise profit by graphical method.

Formulation of LPP

x_1 = No. of units of A/Month

x_2 = No. of units of B/Month

Max $Z = 100\ x_1 + 80\ x_2$

Subject to constraints:

$6\ x_1 + 4\ x_2 \leq 7200$

$2\ x_1 + 4\ x_2 \leq 4000\ x_1, x_2 \geq 0$

Step 2 Determining each axis What goes there? Horizontal (X) axis — Product A (x_1)Vertical (Y) axis — Product B (x_2)

Step 3: Calculation of constraint lines to draw the constraint lines on graphs.
At this point, the constraints are in the form of inequality (≤). So lets make them equal to get co-ordinates.

Constraint No. 1:

$6\ x_1 + 4\ x_2 \leq 7200$

Converting into equality:

$6\ x_1 + 4\ x_2 \leq 7200$

x_1 is the intercept on X axis and x_2 is the intercept on Y axis. To find x_1, let $x_2 = 0$

$6\ x_1 = 7200$

$6\ x_1 = 7200$

$\therefore x_1 = 1200;\ x_2 = 0\ (1200, 0)$

To find x_2, let $x_1 = 0$

$4\ x_2 = 7200$

$\therefore x_2 = 1800;\ x_1 = 0\ (0, 1800)$

Hence the two points which make the constraint line are:
(1200, 0) and (0, 1800)

Note : When we write co-ordinates of any point, we always write (x_1, x_2) You can read that value of x1 followed by value of x2. Therefore, if for a point x_1=1200 and x_2=0 then its co-ordinates will be (1200, 0).
For the second point, x1=0 and x2=1800. Its co-ordinates, therefore, will be (0, 1800)

Constraint No. 2:

$2\ x_1 + 4\ x_2 \leq 4000$

To find x_1, let $x_2 = 0$

$2 x_1 = 4000$
$\therefore x_1 = 2000;\ x_2 = 0$ (2000, 0)
To find x_2, let $x_1 = 0$
$4 x_2 = 4000$
$\therefore x_2 = 1000;\ x_1 = 0$ (0, 1000)
On the graph, each of these constraints will be shown with one straight line. Since there are two constraints, we will get two straight lines.
Co-ordinates of points are:

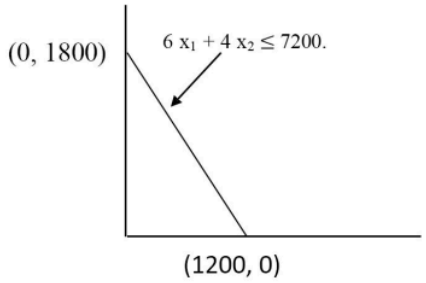
1. Constraint No. 1: (1200, 0) and (0, 1800)
2. Constraint No. 2: (2000, 0) and (0, 1000)

Step 4: Graphing the constraint lines
We choose a suitable scale to mark these points on the graph. Scale to take will depend on maximum value of x1 & x2 from co-ordinates.
For x_1, we have 2 values ⟶ 1200 and 2000
$\therefore$Max. value for $x_2 = 2000$
For x_2, we have 2 values ⟶ 1800 and 1000
$\therefore$Max. value for $x_2 = 1800$

Let us suppose that we have a graph paper of size 20 X 30 cm. Lines need to have some buffer on both axis, so the maximum value of your 20 cm of x-axis is 2000.
$\therefore$ Scale 1 cm = 200 units
$\therefore$ 2000 units = 10 cm (X-axis)
1800 units = 9 cm (Y-axis)
Use a large enough size, so the diagram shoulden't be too little.

Constraint No. 1:
The constraint $6 x1 + 4 x2 \leq 7200$ is represented by the line joining the two points (1200, 0) and (0, 1800).
Fig 1.

Fig 1.

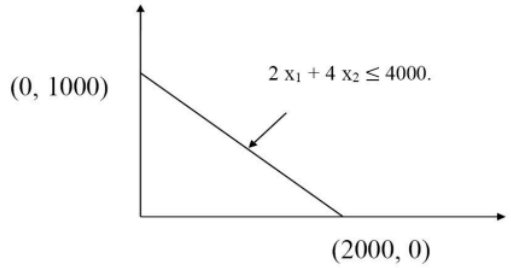


As the equation holds (equality) for all points on the line.
$6 x_1 + 4 x_2 \leq 7200$. This means that every point is below the line will satisfy the inequality (less than): $6 x_1 + 4 x_2 \leq 7200$.

Constraint No. 2:
The line connecting the two points (2000, 0) and (0, 1000) represents constraint $2 x_1 + 4 x_2 \leq 4000$
We know if y is expressed as a function of x, we are given an equation of a line (equality), and all points on the line will satisfy.

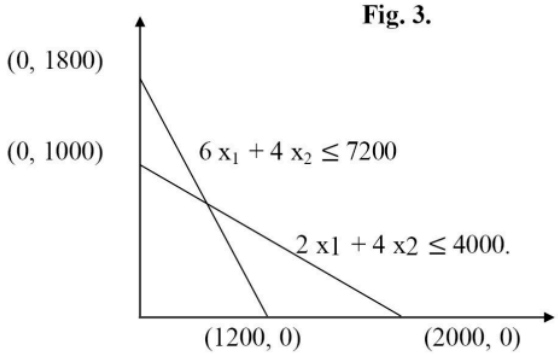
$2x_1 + 4x_2 \leq 4000$
All of the points below the line will be solutions to the inequality (less than).
$2x_1 + 4x_2 \leq 4000$

Fig 2.



So now we will have something like this in our final graph:
Fig. 3.

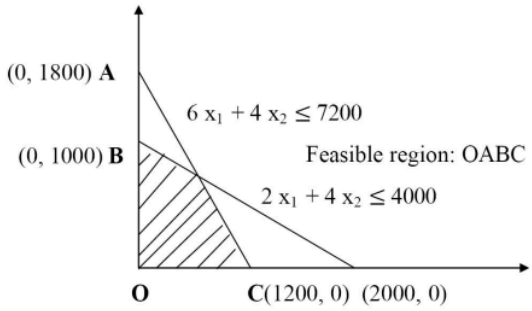
Fig. 3.



Step 5: Determining Feasible Regions
The constraint lines form the boundaries of a feasible region. All points that lie inside the feasible region or on the boundary of the feasible region or on the corner of the feasible region are the points that are satisfiable for all the constraints. Both the constraints are of 'less than or equal to' (≤) type. So the solution region must lie within both constraint lines.

Therefore, the polygon OABC is the feasible region. What is (0, 0)? O is origin (2) O, A, B and C are vertices of the feasible region.

Fig 4. **Scale 1 cm = 200 units**



Step 6: Finding the optimal Solution
The final answer will always be at a point on the feasible area (choose any point on the covered area).
To find optimal solution:
We use corner point method. At each vertex or corner point we identify (x_1, x_2 Values). From this we get 'Z' value for every corner point.

Vertex	Co-ordinates	**Z = 100** x_1 **+ 80** x_2
O	$x_1 = 0, x_2 = 0$ From Graph	Z = 0
A	$x_1 = 0, x_2 = 1000$ From Graph	Z = Rs. 80, 000
B	$x_1 = 800, x_2 = 600$ From Simultaneous equations	Z = Rs. 1, 28, 000
C	$x_1 = 1200, x_2 = 0$ From Graph	Z = Rs. 1, 20, 000

Max. Z = Rs. 1, 28, 000 (At point B)
Solution
Optimal Profit = Max Z = Rs. 1,28,000
Product Mix: x1 = No. of units of A / Month = 800
x2 = No. of units of A / Month = 600

16.2.3 Minimisation Model
Example
A company is involved in the breeding of animals. The animals will receive nutrition supplements daily. There are two products A and B that include the three necessary nutrients.

Nutrients	Quantity/unit		Minimum Requirement
	A	**B**	
1	72	12	216
2	6	24	72
3	40	20	200

Product cost per unit are: A: Rs. 40; B: Rs. 80. Find out quantity of product A & B to be given to provide minimum nutritional requirement.
Step 1: Formulation as LPP
x_1 - Number of units of A
x_2 - Number of units of B
Z - Total Cost
Min. $Z = 40 x_1 + 80 x_2$
Subject to constraints:
$72 x_1 + 12 x_2 \geq 216$
$6 x_1 + 24 x_2 \geq 72$
$40 x_1 + 20 x_2 \geq 200$ $x_1, x_2 \geq 0$.
Step 2: Determination of each axis
Horizontal (X) axis: Product A (x_1)
Vertical (Y) axis: Product B (x_2)

Step 3: Obtain the co-ordinates of the constraint lines to plot the graph. The only types of constraints you have are 'greater than or equal to'. Let's change them into equality:

Constraint No. 1:

$72\ x_1 + 12\ x_2 \geq 216$

Converting into equality

$72\ x_1 + 12\ x_2 = 216$

To find x_1, let $x_2 = 0$

$72\ x_1 = 216$ $\therefore x_1 = 3, x_2 = 0$ (3, 0)

To find x_2, let $x_1 = 0$

$12\ x_2 = 216$ $\therefore x_1 = 0, x_2 = 18$ (0, 18)

Constraint No. 2:

$6\ x_1 + 24\ x_2 \geq 72$

To find x_1, let $x_2 = 0$

$6\ x_1 = 72$ $\therefore x_1 = 12, x_2 = 0$ (12, 0)

To find x_2, let $x_1 = 0$

$24\ x_2 = 72$ $\therefore x_1 = 0, x_2 = 3$ (0, 3)

Constraint No. 3:

$40\ x_1 + 20\ x_2 \geq 200$

To find x_1, let $x_2 = 0$

$40\ x_1 = 200$ $\therefore x_1 = 5, x_2 = 0$ (5, 0)

To find x_2, let $x_1 = 0$

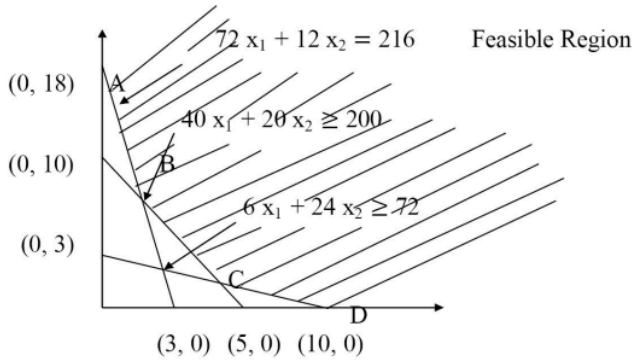
$20\ X2 = 200$ $\therefore x_1 = 0, x_2 = 10$ (0, 10)

The co-ordinates of points are:

1. Constraint No. 1: (3, 0) & (0, 18)
2. Constraint No. 2: (12, 0) & (0, 3)
3. Constraint No. 3: (5, 0) & (0, 5)

And each point with this equation will satisfy (equation equality)
72 x1 + 12 x2 = 216.
(>) is greater than the line.
72 x1 + 12 x2 = 216.
The same we can do for other two constraints, and we can graph lines.

Step 5: Feasible Region



All constraints are of greater than or equal to (≥) type.
That's why feasible region should be above (to the right of) all constraints.
The extreme Points of the feasible region are A, B, C & D.

Step 6: Solve for the optimal solution
Corner Point Method

Vertex	Co-ordinates	$Z = 40\ x_1 + 80\ x_2$
A	$x_1 = 0, x_2 = 18$ From Graph	$\therefore Z = 1,440$
B	$x_1 = 2, x_2 = 6$ From Simultaneous Equations	$\therefore Z = 560$
C	$x_1 = 4, x_2 = 2$ From Simultaneous Equations	$\therefore Z = 320$
D	$x_1 = 12, x_2 = 0$ From graph	$\therefore Z = 480$

We find from table Min. Z = Rs. 320 (At point 'C')
Optimal Product Mix:
$x_1 = 4$
$x_2 = 2$
Minimum value of Z = Rs. 320

16.2.4 Maximisation-Mixed Constraints

Example

A company produces two products P1 and P2 and has a production capacity of 18 tons per day. P1 & P2 have different production capacity. Per day, the firm must provide at least 4 t of P1 & 6 t of P2. P1 and P2 – require 60 hours of machine work each per tonne. The maximum machine hours available are 720. P1's profit per tonne is Rs160, and P2's is Rs240. Using graphical method to solve linear programming problem

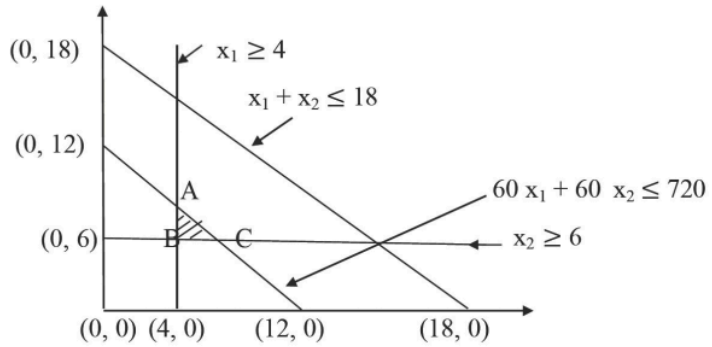
LPP Formulation

x_1 = Tonnes of P1 / Day
x_2 = Tonnes of P2 / Day
Max. $Z = 160\ x_1 + 240\ x_2$
Subject to constraints
$x_1 \geq 4$
$x_2 \geq 6$
$x_1 + x_2 \leq 18$
$60\ x_1 + 60\ x_2 = 0$
For the lines that describe the constraints:

1. $x_1 \geq 4$(4, 0) — No value for x_2, $\therefore x_2 = 0$
2. $x_2 \geq 6$(0, 6) — No value for x_1, $\therefore x_1 = 0$
3. $x_1 + x_2 \leq 18$ — (18, 0) (0, 18)
4. $60\ x_1 + 60\ x_2 \leq 720$ — (12, 0) (0, 12)

If $x_1 = 0$, $60\ x_2 = 720 \therefore x_2 = 12$(0, 12)
If $x_2 = 0$, $60\ x_1 = 720 \therefore x_1 = 12$ (12, 0)
Graph:
x_1: X Axis
x_2: Y Axis
Scale:

Maximum value for $x_1 = 18$;
Maximum value for $x_2 = 18$; ∴
Scale: 1 cm = 2 Tonnes.



Two of the constraints are of the type 'greater than or equal to'. So above or right from these constraint lines will the feasible region lie. And two constraints are of type 'less than or equal to'. So feasible region would be below or left of these constraint line. Therefore, feasible region is ABC.
Perfect Solution Taper Corner Point Strategy

Vertex	Coordinates	$Z = 160\ x_1 + 240\ x_2$
A	$x_1 = 4, x_2 = 8$ Simultaneous Equation	∴ Z = Rs. 2, 560
B	$x_1 = 4, x_2 = 6$ From Graph	∴ Z = Rs. 2, 080
C	$x_1 = 6, x_2 = 6$ Simultaneous Equations	∴ Z = Rs. 2, 400

From Table we find that Maz Z = 560 at Point A
$x_1 = 4$
$x_2 = 8$
Optimal Profit Z. Max = Rs. 2, 560.

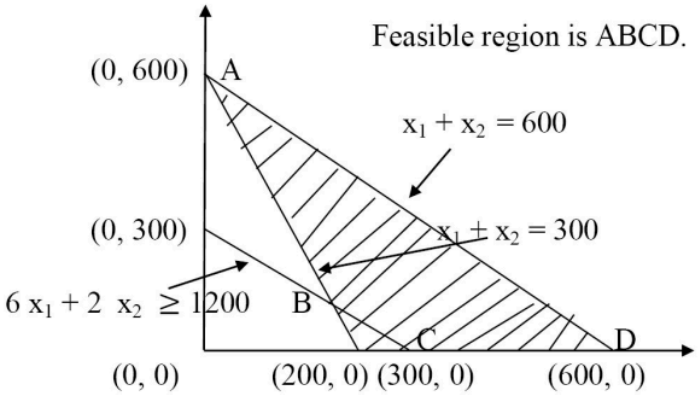
16.2.5 Minimisation Mixed Constraints

Example 1:

A company manufactures two products P and Q. It has an upper limit of 600 units of total production per day. However, each day at least 300 total units must be made. Machine hours per unit: 6 for P and 2 for Q Machine hours must be used at least 1200 machine hours must daily. P has a manufacturing cost of Rs. 50 per unit, and Q has a Rs. 20 cost. Graphically determine the LPP's optimal solution.
LPP formulation
x_1 = No. of Units of P / Day
x_2 = No. of Units of Q / Day
Min. $Z = 50\ x_1 + 20\ x_2$
Constraints $x_1 + x_2 \leq 600$
$x_1 + x_2 \geq 300$
$6\ x_1 + 2\ x_2 \geq 1200$
$x_1, x_2 \geq 0$
Coordinates for Constraint lines

1. $x_1 + x_2 = 600$
   If $x_1 = 0$, $x_2 = 600$ $\therefore$ (0, 600)
   If $x_2 = 0$, $60x_1 = 600$ $\therefore$ (600, 0)
2. $x_1 + x_2 = 300$
   If $x_1 = 0$, $x_2 = 300$ $\therefore$ (0, 300)
   If $x_2 = 0$, $60\ x_1 = 300$ $\therefore$ (300, 0)
3. $6\ x_1 + 2\ x_2 \geq 1200$
   If $x_1 = 0$, $2x_2 = 1200$ $\therefore\ x_2 = 600$ (0, 600)
   If $x_2 = 0$, $6\ x_1 = 1200$ $\therefore\ x_1 = 200$ (200, 0)

Graph:
x_1: X Axis
x_2: Y Axis
Maximum value for $x_1 = 600$;
Maximum value for $x_2 = 600$;
$\therefore$ Scale: 1 cm = 50 units.



In total, there are two {'greater than or equal to'} type constraints. Thus, for constraint lines, feasible region will be greater than or equal to these lines. Two of the constraints are of the 'less than or equal to' variety. Therefore feasible region will lie below or left to these constraints line. Therefore feasible region is ABCD.

Optimal Solution Corner Point Method

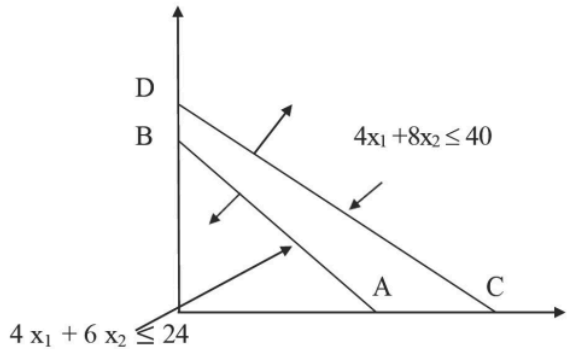
Vertex	Coordinates	$Z = 160\ x_1 + 240\ x_2$
A	$x_1 = 0$, $x_2 = 600$ From Graph	$\therefore$ Z = Rs. 12, 000
B	$x_1 = 150$, $x_2 = 150$ Simultaneous Equations	$\therefore$ Z = Rs. 10, 500
C	$x_1 = 300$, $x_2 = 0$ From Graph	$\therefore$ Z = Rs. 15, 000
D	$x_1 = 600$, $x_2 = 0$ From Graph	$\therefore$ Z = Rs. 30, 000

Min. Z = Rs. 10, 500

Solution
Optimal Cost = Rs. 10, 500/-
x_1 = No. of Units of P = 150
x_2 = No. of Units of P = 150.

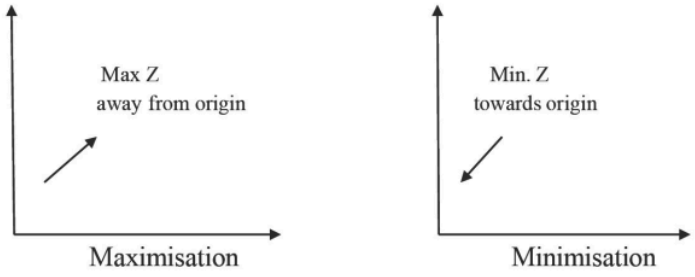
16.2.6 Linear Programming : Special Cases
a. No Solution (Infeasibility)
Infeasible definition: Not possible. The infeasible solution occurs when the constraints are contradictory to each other. No solution can be found that meets all constraints. Feasibility using graphical method In the graphical method, infeasibility arises when we cannot find Feasible region.

Example 1:
Max. Z = 5 x1 + 8 x2
Subject to constraints
4 x1 + 6 x2 ≤ 24
4 x1 + 8 x2 ≤ 40
x1, x2 ≥ 0

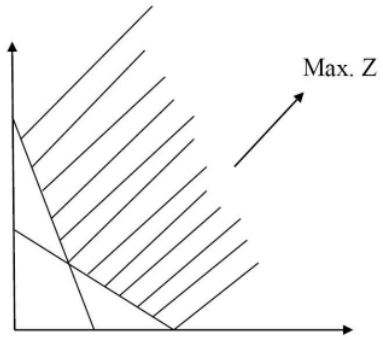


"Because AB and CD have no common feasible region. Thus, solution is not possible.

b. Rescheduling Solution (Overconstrainedness)
Solution mean infinite for Unbounded. Unbounded solution: If a solution has infinity answer, it is called unbounded solution.
In graphical solutions, the direction w.r.t origin is:



Now if we have following feasible region in a maximisation problem:



Since there is no upper bound (away from origin), thus the answer is infinity. This type is known as unbounded solution.

c. Redundancy (Redundant Constraint)

Redundant means that the constraint does not belong to our solution. That constraint has no effect on the feasible region.

This implies that even if we had removed the restriction from our solution we wouldn't have changed our optimal solution.

Example

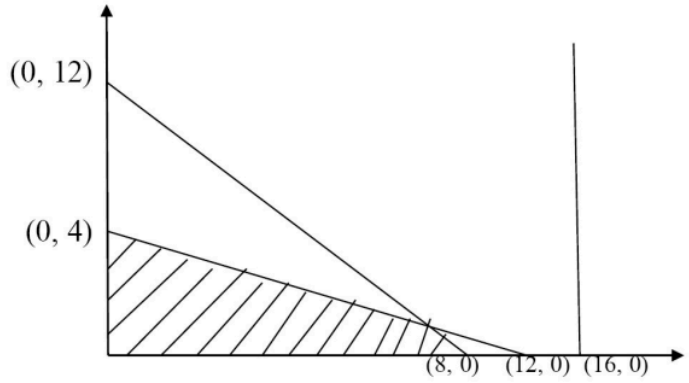
Max. $Z = 5\ x1 + 8\ x2$

Subject to Constraints

$3\ x1 + 2\ x2 \leq 24$

$x1 + 3\ x2 \leq 12$

$x1 \leq 16\ x1, x2 \geq 0$



The feasible region for the above problem is OABC. The 3rd constraint is redundant and doesn't affect the feasible region.

This is why the constraint $x1 \leq 16$ is said to be a redundant constraint.

d. Alternate Optimal Solution: (Multiple Optimal Solution)

Alternate/multiple optimal solution: It means a problem has one or more than one solution which gives the optimal solution.

Multiple sets of solution values is available which provides the maximum profit or minimum cost. When using graphical method, we find out that there exist alternative optimal solution, when:

The iso-cost (or iso-profit) line will be parallel to one of the sides of the feasible region (they will have the same value of slope).

16.2.7 Exercises

1. On what is feasible region in graphical method
2. Explain 'iso-profit' and 'iso-cost line' in graphical solution.
3. A.P. Ravi wants to invest Rs. 1, 00, 000 in two companies 'A' and 'B' in such a way that he does not invest more than Rs. 75, 000 in either of the company. At company 'A' you are guaranteed return of 10% wheres at company 'B' you get 20% average return. For example, the risk factor rating for company 'A' is 4, on a scale of zero to ten while that for company 'B' is 9 on the same scale. Since Mr. Ravi hopes to optimise his returns, he will not accept risk-adjusted average rate of return below 12% risk or risk factor above 6. Assume data to October 2023 and solve graphically by formulating this as LPP
4. Graphically solve the following LPP and interpret the solution.
   Max. $Z = 8x_1 + 16 x_2$
   Subject to:
   $x_1 + x_2 \leq 200$
   $x_2 \leq 125$
   $3x_1 + 6x_2 \leq 900$
   $x_1, x_2 \geq 0$
5. There is a furniture manufacturer who produces two products – tables → chairs. These products are processed on two types of machines A and B. For processing a chair, it requires 2 hours on machine type A and 6 hours on machine type B; while a table requires 5 hours on machine type A and zero on Machine type B. There are 16 hours/day available on machine type A and 30 hours/day on machine type B. The profits gained by the manufacturer from a chair and a table are Rs. 2 and Rs. 10 respectively. Daily production of each of the two products should be? Get solution using LPP graphical method.

16.3 SUMMARY:

- Linear programming determines the way to achieve the best outcome (such as maximum profit or lowest cost) in a given mathematical model and some list of requirements represented as linear equations.
- It is a technique to ensure the optimum allocation of scarce resources in order to deliver for the fulfillment of ever increasing demands in the market.
- Linear Programming is used as a helping tool in nearly all functional areas of management.
- The graphical method to solve linear programming problem helps to visualize the procedure explicitly.
- It also helps to understand the different terminologies associated with the solution of LPP.

16.4 TECHNICAL TERMS:

- **Constraints:** A condition that a solution to an optimization problem must satisfy.
- **Feasible Region:** The region containing solution.
- **Feasible Solution:** If a solution satisfies all the constraints, it is called feasible solution.

16.5 REFERENCES:

1. Sharma, A. (2009). Operations Research. Global Media, Himalaya Publishing House.
2. Sharma, J.K. (2010), Operations Research – Problems and Solutions, Third Edition, Macmillan Publishers India Ltd.
3. Taha, H. A. (2008), Operations Research – An Introduction, Eight Edition, Prentice – Hall of India Private Ltd.

Dr Pachala Vijaya Vani

LESSON- 17

SIMPLEX METHOD

OBJECTIVES:

The purpose of studying this chapter is :

- Describe the principle of simplex method
- Discuss the simplex computation

STRUCTURE:

17.1 INTRODUCTION:

The graphical technique of handling linear programming problems is useful for understanding its fundamental structure, but industrial problems include many variables, making it difficult. Simplex Method solves linear programming problems with more variables. The strategy iteratively functions and obtains the objective function's maximum or minimum value. This approach helps decision makers find redundant constraints, unbounded solutions, numerous solutions, and infeasible problems.

17.2 PRINCIPLE OF SIMPLEX METHOD:

The methodological characteristics of the Simplex technique are presented using a linear programming issue involving two choice variables in the subsequent section.

Example: Maximize $50x_1 + 60x_2$

Subject to :

$$2x_1 + x_2 \leq 300$$
$$3x_1 + 4x_2 \leq 509$$
$$4x_1 + 7x_2 \leq 812$$
$$x_1 \geq 0, \quad x_2 \geq 0$$

Solution

We define the variables $x_3 \geq 0$, $x_4 \geq 0$, $x_5 \geq 0$. Thus, the constraints transform into equations.

$$\begin{aligned} 2x_1 + x_2 + x_3 &= 300 \\ 3x_1 + 4x_2 + x_4 &= 509 \\ 4x_1 + 7x_2 + x_5 &= 812 \end{aligned}$$

The variables x_3, x_4, and x_5 are referred to as slack variables associated with the three constraints. The system of equations has five variables, including the slack variables, and consists of three equations.

17.2.1 Basic Solution

In the above mentioned system of equations, we can set any two variables to zero. The system has three equations with three variables. A solution to this system of three equations with three variables, if it exists, is referred to as a **basic solution.**

In the aforementioned case, let us assume $x_1 = 0$ and $x_2 = 0$. The solution for the system with the remaining three variables is $x_3 = 300$, $x_4 = 509$, $x_5 = 812$. This is a **basic solution** of the system. The variables x_3, x_4, and x_5 are called as **basic variables**, where the variables x_1 and x_2 are designated as non-basic variables, which are set to zero.

17.2.2 Basic Feasible Solution

A basic solution of a linear programming problem qualifies as a **basic feasible solution** if it is feasible, meaning all variables are non-negative. The answer $x_3 = 300$, $x_4 = 509$, $x_5 = 812$ constitutes a **basic feasible solution** to the problem.

In the presence of several variables in a linear programming problem, it is not feasible to geometrically ascertain the extreme points. However, we can discern them through the basic feasible solutions. As one of the basic feasible solutions would optimize the objective function, we can begin this search from one basic feasible solution to another. The simplex approach systematically searches for an optimal solution, continuously increasing the objective function (in maximization case) until the fundamental feasible solution is identified at which the objective function reaches its maximum value.

17.2.3 The slack, surplus & artificial variables:

In case of the inequality be $\leq$ (less than or equal, then we append a slack variable + S $\rightarrow$ and convert $\leq$

If the inequality be $\geq$ (greater than or equal, then we subtract a surplus variable - S $\rightarrow$ and convert $\geq$.

And in both the case to make = artificial variables are added.

17.3 COMPUTATIONAL ASPECT OF SIMPLEX METHOD:

Step 1: Find an initial basic feasible solution.

Step 2: Use the optimality condition to select an entering variable If there is no entering variable, stop.

Step 3: Apply the feasibility condition to select a leaving variable.

17.3.1 Optimality condition:
In a maximization (minimization) problem, the entering variable is the non-basic variable with the most negative (positive) coefficient in the Z-row. The optimum condition is obtained at the iteration, where all the Z-row coefficient of non-basic variables are either non-negative (or) non-positive.

17.3.2 Feasibility condition:
In case of maximization as well minimization problem the minimum non negative ratio (with strict positive denominator) identifies the leaving variable.

17.3.3 Pivot row:

a) Replacing the leaving variable in the basic column with the entering variable.
b) For the new pivot row = current pivot row/pivot element.
c) All other rows: new row = current row − (pivot column coefficient) ∗ new pivot row

17.4 SIMPLEX METHOD WITH SEVERAL DECISION VARIABLES:

The computational method described in the preceding section may be easily adapted to linear programming problems involving several choice variables. Here are some examples

Example 1: Use the simplex method to determine the (LP) model:
Max $Z = 5x_1 + 4x_2$
Subject to

$$\begin{aligned} 6x_1 + 4x_2 &\le 24 \\ x_1 + 2x_2 &\le 6 \\ -x_1 + x_2 &\le 1 \\ x_2 &\le 2 \\ \text{and} \quad x_1, x_2 &\ge 0 \end{aligned}$$

Solution:
Max $Z = 5x_1 + 4x_2$
Subject to

$$\begin{aligned} 6x_1 + 4x_2 + x_3 &= 24 \\ x_1 + 2x_2 + x_4 &= 6 \\ -x_1 + x_2 + x_5 &= 1 \\ x_2 + x_6 &= 2 \\ x_1, x_2, x_3, x_4, x_5, x_6 &\ge 0 \end{aligned}$$

Table 1:

Basic	x_1	x_2	x_3	x_4	x_5	x_6	Sol.
x_3	6	4	1	0	0	0	24
x_4	1	2	0	1	0	0	6
x_5	-1	1	0	0	1	0	1
x_6	0	1	0	0	0	1	2
Max Z	-5	-4	0	0	0	0	0

Choose the Minimum element of Max Z row (not zero), and
Select the corresponding column where Minimum Element in Max Z row is identified
Divide the expected solution column (last column) with the Corresponding elements of the selected column (entering variable column) i.e and select the minimum element greater than zero
Here we choose -5 corresponding to x_1 column

24/6 = 4
6/1 = 6
1/-1 = -1 (ignore)
2/0 = ∞ (ignore)

24/6 = 4 corresponds to x_3 row

Hence x_1 enters the basis and x_3 leaves the basis

Therefore the common element of entering variable x_1 column and leaving variable x_3 row is 6 which is called as key element or pivot element.
Now divide the pivot element row with pivot element to formulate the next table
New pivot row = current pivot row / pivot element

Now new x_1 row = [6 4 1 0 0 0 24] / 6

= [1 2/3 1/6 0 0 0 4]

New row = current row − (pivot column coefficient) ∗ new pivot row

New x_4 row = [1 2 0 1 0 0 6] – (1) [1 2/3 1/6 0 0 0 4]

= [0 4/3 -1/6 1 0 0 2]

New x_5 row = [-1 1 0 0 1 0 1] – (-1)[1 2/3 1/6 0 0 0 4]

= [0 5/3 1/6 0 1 0 5]

New x_6 row = [0 1 0 0 0 1 2] – (0)[1 2/3 1/6 0 0 0 4]

= [0 1 0 0 0 1 2]

New Z row = [-5 -4 0 0 0 0 0] – (-5)[1 2/3 1/6 0 0 0 4]

= [0 -2/3 5/6 0 0 0 20]

Table 2:

Basic	x_1	x_2	x_3	x_4	x_5	x_6	Sol.
x_1	1	2/3	1/6	0	0	0	4
x_4	0	4/3	-1/6	1	0	0	2

x_5	0	5/3	1/6	0	1	0	5
x_6	0	1	0	0	0	1	2
Max Z	0	-2/3	5/6	0	0	0	20

Choose the Minimum element of Max Z row (not zero), and
Select the corresponding column where Minimum Element in Max Z row is identified
Divide the expected solution column (last column) with the Corresponding elements of the selected column (entering variable column) i.e and select the minimum element greater than zero
Here we choose -2/3 corresponding to x_3 column

4/(2/3) = 6
2/(4/3) = 3/2
5/(5/3) = 3
2/1 = 2

2/(4/3) = 3/2 corresponds to x_4 row

Hence x_2 enters the basis and x_4 leaves the basis

Therefore the common element of entering variable x_2 column and leaving variable x_4 row is 4/3 which is called as key element or pivot element.
Now divide the pivot element row with pivot element to formulate the next table
New pivot row = current pivot row / pivot element

Now new x_2 row = [0 4/3 -1/6 1 0 0 2] / (4/3)

= [0 1 -1/8 3/4 0 0 3/2]

New row = current row − (pivot column coefficient) ∗ new pivot row

New x_1 row = [1 2/3 1/6 0 0 0 4] – (2/3) [0 1 -1/8 3/4 0 0 3/2]

= [1 0 1/4 -1/2 0 0 3]

New x_5 row = [0 5/3 1/6 0 1 0 5] – (5/3) [0 1 -1/8 3/4 0 0 3/2]

= [0 0 3/8 -5/4 1 0 5/2]

New x_6 row = [0 1 0 0 0 1 2] – (1) [0 1 -1/8 3/4 0 0 3/2]

= [0 0 1/8 -34 0 1 1/2]

New Z row = [0 -2/3 5/6 0 0 0 20] – (-2/3) [0 1 -1/8 3/4 0 0 3/2]

= [0 0 5/6 1/2 0 0 21]

Table 3: (optimal solution):

Basic	x_1	x_2	x_3	x_4	x_5	x_6	Sol.
x_1	1	0	1/4	-1/2	0	0	3
x_2	0	1	-1/8	3/4	0	0	3/2
x_3	0	0	3/8	-5/4	1	0	5/2
x_4	0	0	1/8	-3/4	0	1	1/2
Max Z	0	0	5/6	1/2	0	0	21

Since all the elements in Maz Z row are ≥ 0 the solution is optimal

The optimal solution is Z = 21 and $x_1 = 3$, $x_2 = 3/2$

Example 2: Solve the (LP) model using the simplex method:

Min $Z = -6x_1 - 10x_2 - 4x_3$

Subject to

$x_1 + x_2 + x_3 \leq 1000$

$x_1 + x_2 \leq 500$

$x_1 + 2x_2 \leq 700$

and $x_1, x_2, x_3 \geq 0$

Solution:

Min $Z + 6x_1 + 10x_2 + 4x_3 = 0$

Subject to

$x_1 + x_2 + x_3 + x_4 = 1000$

$x_1 + x_2 + x_5 = 500$

$x_1 + 2x_2 + x_6 = 700$

$x_1, x_2, x3, x_4, x_5 x_6 \geq 0$

Table 1:

Basic	x_1	x_2 ↓	x_3	x_4	x_5	x_6	Sol.
x_4	1	1	1	1	0	0	1000
x_5	1	1	0	0	1	0	500
← x_6	1	2 □	0	0	0	1	700
Max Z	6	10 ◯	4	0	0	0	0

Choose the Maximum element of Min Z row (not zero), and

Select the corresponding column where Maximum Element in Max Z row is identified

Divide the expected solution column (last column) with the Corresponding elements of the selected column (entering variable column) i.e and select the minimum element greater than zero

We Choose here 10 corresponding to x_2 coloumn

1000/1 = 1000

500/1 = 500

700/2 = 350

700/2 = 350 corresponds to x_6 row

Hence x_2 enters the basis and x_6 leaves the basis

Therefore the common element of entering variable x_2 column and leaving variable x_6 row is 2 which is called as key element or pivot element.
Now divide the pivot element row with pivot element to formulate the next table
New pivot row = current pivot row / pivot element

Now new x_2 row = [1 2 0 0 0 1 700] / 2

= [1/2 1 0 0 0 1/2 350]

New row = current row − (pivot column coefficient) ∗ new pivot row

New x_4 row = [1 1 1 1 0 0 1000] – (1) [1/2 1 0 0 0 1/2 350]

= [1/2 0 1 1 0 -1/2 650]

New x_5 row = [1 1 0 0 1 0 500] – (1) [1/2 1 0 0 0 1/2 350]

= [1/2 0 0 0 1 -1/2 150]

New Z row = [6 10 4 0 0 0 0] – (10) [1/2 1 0 0 0 1/2 350]

= [1 0 4 0 0 -5 -3500]

Table 2:

Basic	x_1	x_2	x_3	x_1	x_2	x_3	Sol.
x_4	1/2	0	1	1	0	-1/2	650
x_5	1/2	0	0	0	1	-1/2	150
x_2	1/2	1	0	0	0	1/2	350
Max Z	1	0	4	0	0	-5	-3500

Choose the Maximum element of Min Z row (not zero), and
Select the corresponding column where Maximum Element in Max Z row is identified
Divide the expected solution column (last column) with the Corresponding elements of the selected column (entering variable column) i.e and select the minimum element greater than zero
We Choose here 4 corresponding to x_3 coloumn

650/1 = 650
150/0 = ∞ (ignore)
700/2 = ∞ (ignore)

650/1 = 650 corresponds to x_4 row

Hence x_3 enters the basis and x_4 leaves the basis

Therefore the common element of entering variable x_3 column and leaving variable x_4 row is 1 which is called as key element or pivot element.
Now divide the pivot element row with pivot element to formulate the next table
New pivot row = current pivot row / pivot element

Now new x_3 row = [1/2 0 1 1 0 -1/2 650] / 1

= [1/2 0 1 1 0 -1/2 650]

New row = current row − (pivot column coefficient) ∗ new pivot row

New x_5 row = [1/2 0 0 0 1 -1/2 150] – (0) [1/2 0 1 1 0 -1/2 650]

= [1/2 0 0 0 1 -1/2 150]

New x_2 row = [1/2 1 0 0 0 1/2 350] – (0) [1/2 0 1 1 0 -1/2 650]

= [1/2 1 0 0 0 1/2 350]

New Z row = [1 0 4 0 0 -5 -3500] – (4) [1/2 0 1 1 0 -1/2 650]

= [-1 0 0 -4 0 -3 -6100]

Table 3: (optimal solution):

Basic	x_1	x_2	x_3	x_1	x_2	x_3	Sol.
x_3	1/2	0	1	1	0	-1/2	650
x_5	1/2	0	0	0	1	-1/2	150
x_2	1/2	1	0	0	0	1/2	350
Max Z	-1	0	0	-4	0	-3	-6100

Since all the elements in Maz Z row are ≤ 0 the solution is optimal

The optimal solution is Min Z = - (- Max Z) = - 6100 and $x_1 = 0$, $x_2 = 350, x_3 = 650$

17.5 SUMMARY:

The simplex approach is the suitable technique for resolving a linear programming problem including several choice variables. Slack variables are provided to convert inequalities into equations with less than or equal to type restrictions. A certain category of solution referred to as a basic feasible solution is crucial for simplex computing. Every basic feasible solution constitutes an extreme point of the convex set of feasible solutions, and conversely.

A basic feasible solution for a system with m equations and n variables consists of m non-negative variables, referred to as basic variables, and n-m variables, valued at zero, known as non-basic variables. The introduction of slack variables facilitates the identification of a basic feasible solution. The goal function attains its maximum or minimum value at one of the basic possible solutions.

The simplex technique starts with the first basic feasible solution derived from the slack variables, progressively enhancing the objective function's value by introducing a new basic variable while rendering one of the existing basic variables non-basic. The selection of a new basic variable and the exclusion of an existing basic variable are conducted according to certain guidelines to enhance the objective function's value in the updated basic feasible solution.

The iterative process ceases when it is no longer feasible to achieve an improved value of the goal function compared to the current one. The current basic feasible solution is the optimal solution that either maximizes or minimizes the objective function, depending on the context.

17.6 TECHNICAL TERMS:

- A **Slack Variable** corresponding to a less than or equal to type constraint is a non negative variable introduced to convert the constraint into an equation.
- A **Basic Solution** of a system of m equations and n variables (m < n) is a solution where at least n-m variables are zero.
- A **Basic Feasible Solution** of a system of m equations and n variables (m < n) is a solution where m variables are non negative and n-m variables are zero.
- A **Basic Variable** of a basic feasible solution has a non negative value.
- A **Non Basic Variable** of a basic feasible solution has a value equal to zero.
- A **Surplus Variable** corresponding to a greater than or equal to type constraint is a non negative variable introduced to convert the constraint into an equation.
- The **Optimum Solution** of a linear programming problem is the solution where the objective function is maximised or minimised.

17.7 SELF ASSESMENT PROBLEMS:

1) A manufacturer has production facilities for assembling two different types of television sets. These facilities can be used to assemble both black and white and coloured sets. At the present time the firm is producing only one model of each type of set. The black and white set contributes Rs. 150 towards profit while a coloured set contributes Rs. 450 towards profit. The number of coloured television sets manufactured everyday cannot exceed 50 as the number of coloured picture tubes available everyday is 50. Each black and white set requires 6 man-hours of chassis assembly time, whereas each coloured set requires 18 man hours. The daily available man hours for the chassis assembly line is 1800. A black and white set must spend one man hour on. the set assembly line whereas a coloured set must spend 1.6 man hours on the set assembly line. The daily available man hours on this line is 240. A black and white television set requires 0.5 man hours of testing- and final inspection whereas a coloured set requires 2 man hours. The total available man hours per day for testing and inspection is 162.
   Formulate and solve the problem using simplex method so that the profit is maximised.

2) A small scale unit is in a position to manufacture three products A, B and C. Raw material required per piece of product A, B and C is 2 kg, 1 kg, and 2 kg respectively while the total daily availability is 50 kg. The raw material is processed on a machine by the labour force and on a day the availability of machine hours is 30 while the availability of labour hour is 26. The time required per unit production of the three products is given below

Product	Machine Hour	Labour Hour
A	½	1
B	3	2
C	1	1

The net per unit profit from products A, B, C respectively are Rs. 25, Rs. 30 and Rs. 40. Find the linear programming formulation of the problem. Solve the problem by simplex method to obtain the maximum profit per day.

3) Solve the following linear programming problem and give your comments

Maximise $6x_1 + 2x_2 + 4x_3$

Subject to :

$$2x_1 + 3x_2 + x_3 \leq 28$$
$$3x_1 + x_2 + 2x_3 \leq 24$$
$$x_1 + 2x_2 + 3x_3 \leq 35$$
$$x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0$$

4) Solve the following linear programming problem and give your comments

Min $Z = 2x_1 + 3x_2 + x_3$

Subject to

$$3x_1 + 2x_2 + x_3 \leq 3$$
$$2x_1 + x_2 + x_3 \leq 2$$
$$x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0$$

17.8 REFERENCES:

1. Sharma, A. (2009). Operations Research. Global Media, Himalaya Publishing House.
2. Sharma, J.K. (2010), Operations Research – Problems and Solutions, Third Edition, Macmillan Publishers India Ltd.
3. Taha, H. A. (2008), Operations Research – An Introduction, Eight Edition, Prentice – Hall of India Private Ltd.

Dr Pachala Vijaya Vani

CHAPTER- 18
BIG M METHOD

OBJECTIVES:

The purpose of studying this chapter is :

- Describe the principle of Big M Method
- Discuss the Mixed Inequalities
- Discuss the Computations with Surplus & Artificial Variables

STRUCTURE:

18.1 INTRODUCTION:

In the preceding instances, we have examined problems that, in normal linear programming form, had a well-defined initial solution with all slack variables. When all constraints in the issue are inequalities, an all-slack solution is the sole viable solution. Now we will examine methodologies for linear programs with alternative forms of constraints.

Note that simplex requires a beginning point. It should transition from one fundamental feasible solution to another, resulting in an improved objective value. An initial basic feasible solution or dictionary can be obtained by designating all slack variables as basic and all original variables as non-basic. Clearly, these assumptions do not hold true for all LPs.

18.2 PRINCIPLE OF BIG M METHOD:

Step 1- Interpret the problem in the standard form

Step 2- Add non- negative artificial variable on the left hand side of each of the equations for constraints of the '$\geq$' or ' =' type.
The introduction of artificial variables makes the corresponding constraints infeasible. We overcome this by marking the artificial variables as zero in the final solution provide the solution exists.
If the problem is infeasible, at least one the artificial variables will have a positive value in the final solution. In the objective function, we assign these variables a very large price (per unit penalty). Such big will be referred to as –M for maximal problems (+M for minimal problem), with $M > 0$.

Step 3 - Finally, we make use of the artificial variables for the initial solution and we continue with the simplex procedure to obtain a solution.

18.3 COMPUTATIONAL ASPECT OF BIG M METHOD:

Example 1 : Find the optimal solution for the following LPP.

Max $Z = 5x_1+12x_2+4x_3$

Subject to the Constraints

$$x_1+2x_2+x_3 \leq 5$$

$$2x_1+x_2+3x_3 = 2$$

$$x_1 \geq 0, x_2 \geq 0 , x_3 \geq 0$$

Solution : Convert Z to the standard form:

$$Z = 5x_1 + 12x_2 + 4x_3 - MR$$

Subject to the Constraints

$$x_1+2x_2+x_3+S_1 = 5$$

$$2x_1-x_2+3x_3+R = 2$$

Implies $R= 2- 2x_1+x_2-3x_3$

Substituting in Z

$$Z = 5x_1+12x_2+4x_3 - M(2-2x_1+x_2-3x_3)$$

$$= (5+2M)x_1+(12-M)x_2 + (4+3M)x_3 -2M$$

Implies $Z - (5+2M) x_1 - (12-M)x_2 - (4+3M)x_3 = -2M$

Subject to the Constraints

$$x_1 + 2x_2 +x_3 + S_1 = 5$$

$$2x_1 - x_2+3x_3+ R = 2$$

Basic var.	x_1	x_2	x_3	S_1	R	SOLU
Z	-5-2M	-12+M	-4-3M	0	0	-2M
S_1	1	2	1	1	0	5
R	2	-1	3	0	1	2
Z	-7/3	-40/3	0	0	4/3+M	8/3
S_1	1/3	7/3	0	1	-1/3	13/3
x_3	2/3	-1/3	1	0	1/3	2/3
Z	-3/7	0	0	40/7	-4/7+M	192/7
x_2	1/7	1	0	3/7	-1/7	13/7
x_3	5/7	0	1	1/7	2/7	9/7
Z	0	0	3/5	29/5	-2/5+M	141/5
x_2	0	1	-1/5	2/5	-1/5	3/5
x_1	1	0	7/5	1/5	2/5	9/5

Example 2 : Determine the optimal solution for the following LPP.

Min $Z = 4x_1+x_2$

Subject to $3x_1 + x_2 = 3$

$4x_1 + 3x_2 \geq 6$

$x_1 + 2x_2 \leq 4$

and $x_1 \geq 0$, $x_2 \geq 0$

Solution : We write Z into the standard form

Min $z = 4x_1+x_2+MR_1+MR_2 + S_1$

Subject to $3x_1+ x_2 + R_1 = 3$

$4x_1 + 3x_2 - S_1 + R_2 = 6$

$x_1 + 2x_2 + S_2 = 4$

$x_1, x_2, S1, S2, R_1, R_2 \geq 0$

$3x_1 + x_2 + R_1 = 3$ implies $R_1 = 3 - 3x_1 - x_2$

$4x_1 + 3x_2 - S_1 + R_2 = 6$ implies $R2 = 6 - 4x_1 - 3x_2 + S_1$

Substituting in Z

$Z = 4x_1 + x_2 + M(3-3x_1-x_2) + M(6-4x_1-3x_2+S1)$

$= (4-7M)x_1 + (1-4M)x_2 + MS1 + 9M$

implies $Z - (4-7M)x_1 - (1-4M)x_2 - MS_1 = 9M$

Subject to $3x_1+x_2+R1 = 3$

$4x_1+3x_2-S1+R2 = 6$

$x_1+2x_2+S2 = 4$

Basic var	x_1	x2	S_1	S_2	R_1	R2	Solu.
Z	4-7M	-1+4M	-M	0	0	0	9M
R_1	3	1	0	0	1	0	3
R_2	4	3	-1	0	0	1	6
S_2	1	2	0	1	0	0	4
Z	0	1/3+5/3M	-M	0	4/3-7/3M	0	4+2M
x_1	1	1/3	0	0	1/3	0	1
R_2	0	5/3	-1	0	-4/3	1	2
S_2	0	5/3	0	1	-1/3	0	3
Z	0	0	1/5	0	8/5-M	-1/5-M	18/5

x_1	1	0	1/5	0	3/5	-1/5	3/5
x_2	0	1	-3/5	0	-4/5	3/5	6/5
S_2	0	0	1	1	1	-1	1
Z	0	0	0	-1/5	7/5-M	-M	17/5
x_1	1	0	0	-1/5	2/5	0	2/5
x_2	0	1	0	3/5	-1/5	0	9/5
S_1	0	0	1	1	1	-1	1

Example 3 : Max $Z = 3x_1 + 2x_2 + x_3$

Subject to

$2x_1 + x_2 + x_3 = 12$

$3x_1 + 4x_3 = 11$

$x_2 \geq 0$, $x_3 \geq 0$ and x_1 is unrestricted

Solution : Convert Z to the standard form

Max $z = 3(x_1' - x_1'') + 2x_2 + x_3 - MR1 - MR2$

subject to

$2(x_1' - x_1'') + x_2 + x_3 + R_1 = 12$

$3(x_1' - x_1'') + 4x_3 + R_2 = 11$

and x_1', x_1'', x_2, x_3, R_1, $R_2 \geq 0$

Max $Z = 3x_1' - 3x_1'' + 2x_2 + x_3 - MR_1 - MR_2$

subject to

$2x_1' - 2x_1'' + x_2 + x_3 + R1 = 12$

$3x_1' - 3x_1'' + 4x_3 + R2 = 11$

x_1', x_2'', x_2, x_3, R_1, $R_2 \geq 0$.

Cj	3	-3	2	1	-M	-M	
Basic var	x_1'	x_1''	x_2	x_3	R_1	R_2	SOL
R_1	2	-2	1	1	1	0	12
R_2	3	-3	4	0	0	1	11
Z	-5M-3	5M+3	-5M-2	-M-1	0	0	-23M
R_1	0	0	-5/3	1	1	x	-14/3
x_1'	1	-1	4/3	0	0	x	-11/3
Z	0	0	5/3M+2	-M-1	0	x	-14M/3-11
x_3	0	0	-5/3	1	x	x	14/3
x_1'	1	-1	4/3	0	x	x	11/3
Z	0	0	1/3	0	x	x	47/3

Now as all $x \geq 0$, optimal basic feasible solution is obtained

$x_1' = 11/3$, $x_1''=0$, $x_1=x_1'-x_1'' =11/3 – 0 =11/3$

Hence the Final solution: Max $z = 47/3$, $x_1 =11/3$, $x_2=0$, $x_3=14/3$

18.4 SUMMARY:

In this unit, you learned the mechanics of obtaining an optimal solution to a linear programming problem by the simplex method. The simplex method is an appropriate method for solving a ≤ type linear programming problem with more than two decision variables. Big M method are used to solve problems of ≤ or ≥ type constraints. Further, the simplex method can also identify multiple, unbounded and infeasible problems.

18.5 TECHNICAL TERMS:

- **Artificial Variables**: Temporary slack variables which re used for calculations.
- **Simplex Method**: A method which examines the extreme points in a systematic manner, repeating the same set of steps of the algorithms until an optimal solution is reached.
- **Slack Variables**: Amount of unused resources.
- **Surplus Variables**: A surplus variable represents the amount by which solution exceeds a resource.
- **Unconstrained Variable**: The variable having no non-negativity constraint.

18.6 SELF ASSESMENT PROBLEMS:

1. Minimize : $Z = 3x_1 + 4x_2$
   Subject to

   $2x_1 + x_2 \leq 6$

   $2x_1 + 3x_2 \geq 9$

   with x_1, x_2 non-negative.

2. Minimize: $Z = -x_1 + x_2$
   Subject to

   $x_1 + 2x_2 \geq 5000$

   $5x_1 + 3x_2 \geq 12000$

   with x_1, x_2 non-negative.

18.7 REFERENCES:

1. Sharma, A. (2009). Operations Research. Global Media, Himalaya Publishing House.
2. Sharma, J.K. (2010), Operations Research – Problems and Solutions, Third Edition, Macmillan Publishers India Ltd.
3. Taha, H. A. (2008), Operations Research – An Introduction, Eight Edition, Prentice – Hall of India Private Ltd.

Dr Pachala Vijaya Vani

CHAPTER- 19
SIMULATION

OBJECTIVES:

The purpose of studying this chapter is :

- Discuss the need for simulation -in management problems where it will not be possible to use precise mathematical techniques
- Explain that simulation may be the only method in situations where it will be extremely difficult to observe actual environment
- Describe the process of simulation based on a sound conceptual framework
- Apply simulation techniques in solving queuing and inventory control problems.

STRUCTURE:

19.1 INTRODUCTION:

Analysis by simulation is an exploratory quantitative method that defines a phenomena by the construct of a model of that phenomena and the performing of a set of controllable experiments to forecast a phenomena over time. Watching the experiments is very much like watching the process itself at work. To see how the real process would respond to some variations, we can generate those changes in our model and simulate the response of the real process to them.

For example, if someone is designing an airplane, the designer can solve different equations describing the aerodynamics of the plane. Alternatively, if these equations are too complicated and complex to solve, a scale model can be constructed and tested in a wind tunnel. In simulation, we create mathematical models that we cannot solve, and run them on data samples to "simulate the behaviour" of the system. Therefore, simulation is conducting experiments on the model of a system.

19.2 REASONS FOR USING SIMULATION:

For, of the number of problems, problems, we have S found the (The number n) through straight forward techniques, mathematical solutions to the situation. Examples we can mention are, for instance, the economic order quantity, the simplex solution for a linear programming problem or a branch-and-bound solution for an integer programming problem.

Notably, in each of those cases the complexity of the problem was reduced by some assumptions so that from a mathematical point of view it could be handled. It is not difficult to imagine managerial situations so convoluted that mathematical solution is impossible in any kind of reasonable time period, given the state of the art on mathematics at that time. Simulation is a solid alternative in such cases,

If we are to demand that all managerial problems be solved mathematically, then we will simplify the situation to get to be able to solve the problem; sacrificing realism in order to solve the problem is where we get into serious trouble. While the assumption of normality-in handling a distribution of inventory demand may be an assumption by no means, and frequencies either through accounting principles land precisely average models lead to misperception of orders in Euclidean or yet comparable assets, stocks, and bonds the assumption of linearity in any precise linear programming environment may be wholly unrealistic.

And while in some instances the solutions that result from the simplifications are acceptable to the decision-maker, in other instances they simply are not. In many cases simulation is a suitable alternative to visiting it mathematically.
It does make assumptions, but they're manageable. Simulation is the tool we have to reduce the uncertainty experienced with some management problems when the mathematical evaluation of a model is not possible.

Some reasons why management scientists might opt to use simulation to help solve management problems include:

1. It is hard to see the real environment, estimation is possibly the only way as simulation (It is commonly used in the field of space flight or in the trajectory of satellites.)
2. This is an inscrutable process that cannot be solved mathematically.
3. The real-world way of observing a system may be prohibitively costly. (This makes the operation of a large computer centre under a number of different operating alternatives, too expensive for its viability.)
4. It may not have a chance to run a lot of time. We couldn't wait the required number of years to see the results, for example, if we were studying long-run trends in world population.
5. There is no need to operate the system and watch it unfold, as that can be too disruptive. (If you are deciding how to provide food service in a hospital, the confusion that would ensue from using two different systems long enough to get valid observations might be too great.)

19.3 LIMITATIONS OF' SIMULATION:

Simulation, like everything else, has its one trade- off, and we should be aware of the limitations that the simulation approach presents. These include the facts that

1. Simulation is not precise. It is not optimization and does not provide the answer, it just supplies a suite of the response of the system to eliminate the different operating conditions. This imprecision is hard to quantify in many instances.
2. Good simulation model may-cost you a lot. It can take years in fact to produce a usable corporate planning model.
3. As a result, not every situation can be assessed by simulation; only situations with uncertainty are contenders, and without a stochastic component, all simulated experiments would yield the same answer.
4. No generation of solutions from simulation, only a metric for evaluating solutions. Managers still have to create the solutions they wish to test.

19.4 STEPS IN THE SIMULATION PROCESS:

Every good simulation takes a lot of planning and organization. While simulations can vary in complexity from one situation to the next, in general you will have to go through these steps:

1. Identify the problem or system you wanted to simulate.
2. Come up with the Model you want to use.
3. Test your model, check how the model behaves against how the real problem behaves.
4. Gather the required data to evaluate the model
5. Run the simulation.
6. Examine the output of the simulation and, optionally, modify alternative combinations of inputs to evaluate
7. Run the simulation again with the updated solution.
8. Validate the simulation; this process enhances the probability of any inferences you might make about how the real system would behave by running the simulation being valid.

19.5 SOME PRACTICAL APPLICATIONS OF SIMULATION:

The applications of simulation that have proven return upon investment are far too numerous to mention here. Nevertheless, some illustration of the variety of managerial contexts in which this approach has been employed to assist with the decision process is worthwhile at this point. All the situations that we shall now describe constitute a classical problem area to which simulation can be successfully applied.

19.5.1 The Home Heating Oil Simulation

A petroleum products distribution firm president attended a management seminar on the quantitative techniques. He became interested in the potential role of simulation in testing the relative effectiveness of his eight home-heating-oil delivery trucks' several alternative dispatching methods. His marketing area covered more than three thousand homes with oil tanks of between 55 and 1900 gallons in capacity. His trucks were between 1000 and 5600 gallons, and his bulk plant (the terminal where he kept his heating oil) had a storage capacity of 150,000 gallons. The firm had one transport truck (which used to haul heating oil from the port) but could rent more trucks if it needed to.

The president knew very well that low-temperature periods taxed "his complete delivery system. His eight trucks were unable to keep up with the residential demand, the bulk plant was a source of confusion and inefficiency, and additional transport trucks had to be rented at disadvantageous short-term prices. There are three options, it appears. One was to enhance

truck and bulk plant equipment and personnel so that capacity would be equal to top cold-weather demand. President already knew this was going to be relatively expensive and had already computed that the additional investment in the equipment alone was around Rs 140,000. A second alternative was to ship heating oil to homes more often - that is, to keep customers tanksov half full so demand in times of low-temperature would be reduced. A third option was to swap out all of the small 55 gallon tanks (at company cost) to dramatically leave the paths of delivery trucks, (Fewer stops per day. and fewer gallons delivered per stop would greatly boost the supply of the delivery fleet.) He was also aware of the alternative combinations two and three.

We believe there is no way to solve this problem mathematically (or at least we do not have the mathematical prowess to solve it) so we ought to create a simulation model for this problem which contains these components:

1. The bulk plant
2. The customers
3. Different residential tank locations
4. Local delivery trucks
5. Transportation trucks (owned and hired)
6. Employees
7. Temperature-based heating-oil consumption

I would simulate a number of alternative, delivery, systems across a broad spectrum of demand scenarios. We would then be able to determine the most effective way forward among our options.

19.5.2 The Application What-cuts-the-carpet?

Executive Development Programme A production vice-president bf a regional carpet manufacturing company attended an executive development programme. Then one day he asked us if any of us had ever done any work with "carpet-cutting." And they were soon in the mill watching the whole operation. Carpet was made 175-foot rolls, all 12 feet wide. This place had three hundred different styles and colours of carpets; sometimes there would be two or more rolls or pieces of a roll of each styles and colour in the warehouse. Incoming carpet orders specified lengths from around 8 feet to a full roll (175 feet). Incoming orders were brought to the cutting room, where cut machine operators tried to match those open rolls to incoming orders so that the unusable piece remaining at the end of the roll—the remnant; would be as small as possible. The significance of unusable remnants perhaps can be gauged from the fact that the average price per lineal foot of the carpet was about Rs. 200 and any remnant which was less than 3 feet was discarded; the 3 feet and 6 feet longer remnants were being sold for about a third of the regular price. The cost of non-usable left-overs was close to Rs.2500000 a year.

The operators of the cutting machine explained that there were really hundreds of ways you could fill-an order for a single-carpet: -(1) cut it from the longest roll of the required style and colour available; -(2) cut it from the roll which will leave the shortest piece left over; -(3) find two orders which will together fill a whole roll or (a whole piece of) a roll; -(4) and so on. To make things worse, one of the teachers wanted to know if it would be profitable to collect carpet orders for more than 1 day (2 days, 3 days and so on) before conducting the cut. his theory that you could make better pairs of orders and rolls the more orders you have. You would of course have to be willing to risk the wrath of customers who would be kept waiting longer.

An analysis of some length of the operation may point to a simulator model of the system with such components:

1. The production' operation (how the carpets were supplied to the cutting operation, how often, etc.)
2. The distribution of incoming orders (size, colour, style)
3. The inventory (sizes, colours, styles)
4. The time/employees for cutting process
5. Prices of sold carpets and remains

We should simulate the cutting operations under a high number of different possible "cutting rules", each simulation run must be for at least say 1000 days, a period which seems long enough to restore a typical order and production pattern. Then we may pick that cutting rule which minimized annual expense and has most saving.

19.5.3 A Public Sector Planning Application

A management institute faculty member had been hired to teach a introductory operations research course for senior administrators of one of the countrys largest metropolitan school systems. The superintendent (another course participant), for example, said one day (and then perhaps more than once) that dealing with his school board on some long range planning matters was extremely hard and wondered whether simulation has anything to offer in that context. The board was always asking questions like "what did that mean if enlistments started to grow at a rate of 9% a year instead of 6% each year?" Or "How many years do you think it will be before the population shifts enough to whole series of these mathematically impossible "how, "when" and "what if" questions

One way to address the issue is to recommend a large simulation model of the public school system. It will enable the superintendent to better facilitate long-range planning in what is a very complicated arena. The model has to account for variables like these:

1. Enrolments (by grade, Kindergarten grade 12)
2. Teacher-pupil ratios
3. Classroom capacities
4. Salaries
5. District population
6. Number of panes (with capacities)
7. Number of teachers subject/function/grade wise
8. Construction cost
9. Transportation equipment
10. Warehousing and repair facilities
11. And administrative staff by grade and function
12. Service staff (maintenance, custodial, etc.)

19.6 TWO TYPICAL EXAMPLES OF HAND-COMPUTED SIMULATION:

In this segment, we are going to introduce you to simulation in the context of a problem, which can be simulated manually, that is done without the use of computer.

Example 1: Scheduling patients at a hospital operating room

Table 19.1 : (say) Wednesday Operating Schedule, Room No. 3

Time	Activity	Expected time
8:00 AM	Appendectomy	40 min
8:40	Clean-up	20 min
9:00	Laminectomy	90 min
10:30	Clean-up	20 min
10:50	Kidney removal	120 rain
12:50 PM	Clean-up	20 min
1:10	Hysterectomy	60 min
2:10	Clean-up	20 min
2:30	Colostomy	100 min
4:10	Clean-up	20 min
4:30	Lesion removal	10 min
4:40	Clean-up	20 min

Hospital Simulation

Table 19.1 : Operating Room Number 3 Schedule on Wednesday in a Large Hospital Being able to schedule the last operations on this basis is very difficult less alone the cleaning and disinfection of the operating room, what is apparent from this schedule, and the head operating room nurse thinks it is even impossible to finish with the operating and clean-up - schedule by 5 P.M., the time at which this operating room has to be available for emergency night service.

The hospital management analyst, says, simulation may tell whether the schedule for Wednesday is feasible and if not what adjustment could be made in it. Over the following months, the analyst is able to review the operating room records and determines that patients do not necessarily arrive to the operating room at the time they were scheduled. They frequently must delay the administration of pre-op medication, occasionally it's the people who transport patients to the operating room who fail to appear on time, and from time to time the doctors simply forget to enter an order to take the patient from the floor to the operating room. Analysis Of OR Log Data: Arrival Expectations Table 19.2 The managing analyst discovers that operating time also varies according to the Surgical difficulties encountered, differences in, surgical skills, and the overall effectiveness of the surgical group.

An analysis of operations planned over the previous months produces the results shown in Table 19.3 that provide a very informative indication of this variation. He understands that the deviation of the hxpected clean-up time Simulation will also impact on the time schedule and check the archives again. Here he learns that, half of the time, the clean-up crew clears up in 10 minutes. For the most part, it will take them 30 minutes. With the data collected, he was ready to start the simulation.

Table 19. 2: Arrival Expectations

Patient arrives on time	0.50 probability
Patient arrives 5 minutes early	0.10 probability
Patient arrives i0 minutes early	0.05 probability
Patient arrives 5 minutes late	0.20 probability
Patient arrives 10 minutes late	0.15probability

Table 19.3 : Operation Time Expectations

Operation is completed in the expected time	0.45 probability
Operation is completed in 90% of the expected time	0.15 probability
Operation is completed in 80% of the expected time	0.05 probability
Operation is completed in 110% . of the expected time	0.25 probability
Operation is completed in 120% of the expected time	0.10 probability

Each Process Generator has a "Process" that generates the Variables in the System.
So now the analyst needs to be able to simulate arrival times, operating times, and clean-up times. The process generators are called the methods he uses to dot his. He chooses to use a random number table. (see Appendix). If we do a similar random exercise for a uniformly distributed random variable (whatever that means) (all of its values (in our case digits 0 through 9) are possible in theory) we would expect to see the output as a random number table.

Generating Arrival Times
He uses the first two digits of each 10-digit number in Appendix as the process generator for arrival times. As there are 100 possible two-digit numbers from 00 through 99, he correlates these two digit numbers to arrival variation as follows:

Random numbers		**Arrivals**
00 through 49	On time	(.50 probability)
50 through 59	5 minutes early	(.10 probability)
60 through 64	10 minutes early	(.05 probability)
65 through 84	5 minutes late	(.20 probability)
85 through 99	10 minutes late	(.15 probability)

Generating Operating Times
For operating times, the analyst now uses the last two digits of each 10-digit number in Appendix to generate his process. To return back to where I was, he relates these two digits numbers to operating times like this:

Random numbers		**Operating Times**
00-44	On time completion	(.45 probability)
45-59	Completion in 00% of expected time	(.15 probability)
60-64	Completion in 80% of expected time	(.05 probability)
65-89	Completion in 110% of expected time	(.25 probability)
90-94	Completion in 120% of expected time	(.10 probability)

Generating Clean-up Times
Since the random variable takes on only two values here, he chooses the fourth digit of each 10-digit number in Appendix to be his process generator. So an odd number will stand for a 10-minute clean up, an even number — 30-minute clean-up.

The analyst then continues with the simulation. First, he creates an arrival-time deviation for the first patient: then he creates an operating time deviation for the first operation: then he

generates a clean-up time for that operation. He repeats this process until the last procedure is complete and the theater scrubbed for the last time that day. His simulation results are displayed in Table 19.4.

The analyst simulation suggests the scheduled operations can be done and the mom out by 5 PM. In fact, his simulation shows that the day ends at 4:45 PM, a few minutes before it's supposed to.

Assumptions

The analyst ran a simulation on a day's operations for a single instance, and extrapolating generalizations from such a short simulation might prove to be harmful for us. He could have repeated the day's simulation many times, with different random numbers, and then we could feel better about generalizing from his results. In that assumption, he assumed based on arrival deviation, operating time deviation and clean-up deviation were independent of each other in this simulation. If not, then his simulation is ineffectual. Finally, he applied the discrete distributions of the three variables. If computation time were not an issue, in an ideal world standard continuously distributed random variables would have been suitable.

Table 19.4 : Simulation Result of Activity from Operation Room No. 3

Random number	First two digits	Last two digits	Fourth digit	Meaning	Outcome
15	X			On-time arrival of appendectomy patient	Appendectomy began at 8 A.M.
96		X		Appendectomy completed in 120% of expected time (48 min)	Appendectomy completed at 8:48 Λ.M.
1			X	Clean-up done in 10 min	Room ready for second operation at 8:58 A.M.
9	X			On-time arrival of laminectomy patient (9 A.M.)	Laminectomy began at 9 A.M.
82		X		Laminectomy completed in 110% of expected time (99 min)	Laminectomy completed at 10:39 A.M.
8			X	Clean-up done in 30 min	Room ready for third operation at 11.09 A.M.
41	X			On-time arrival of kidney patient (10:50 A.M.)	Kidney removal began at 11:09. A.M.
56		X		Kidney removal completed in 90% of expected time (108 min)	Kidney removal completed at 12:57 P.M.
2			X	Clean-up done in 30 min	Room ready for fourth operation at 1:27 P.M.
75	X			Hysterectomy patient arrives 5 min late (1:15 P.M.)	Hysterectomy began at 1:27 P.M.
68		X		Hysterectomy completed in 110% of expected time (66	Hysterectomy compicted at 2:33 P.M.

				min)	
Random number	First two digits	Last two digits	Fourth digit	Meaning	Outcome
7			X	Clean-up done in 10 min	Room ready for fifth operation at 2:43 P.M.
00	X			On-time arrival of laminectomy patient (2.30 P.M.)	Colostomy began at 2:43 P.M.
58		X		Laminectomy completed in 90% of expected time (90 min)	Colostomy completed at 4:13 P.M.
9			X	Clean-up done in 10 min	Room ready for sixth operation at 4:23 P.M.
72	X			Lesion patient arrivals 5 min late (4:35 P.M.)	Lesion Operation begun at 4:35 P.M.
40		X		On-time completion of lesion operation (10 min)	Lesion operation completed at 4:45 P.M.
5			X	Clean-up done in 10 min	Operating room schedule for Wednesday completed at 4:55 P.M.

19.6.2 Overview on Simulation and Inventory Management

Info root causes of service pitfalls. to an order at the reorder point must be selected with demand during lead time in mind. However, since the lead time and demand of an inventory per unit time are random variables, the simulation technique can be used to explore the impact of different inventory policies (in other words different order quantity-reorder point combinations) on a probabilistic inventory system.

Example 2: The wholesaler trading with stationary wants to find the desk calendar order size. The demand and lead time (the time it takes for an order to be delivered) are probabilistic and described by their distributions below:,
Order cost per order, Rs. 50 and holding cost for 1000 calendars Rs. 2/wk.

Demand/week (thousand)	Probability	Lead Time (weeks)	probability
0	0.2	2	0.3
1	0.4	3	0.4
2	0.3	4	0.3
3	0.1		

The cost due to shortage is Rs. 10 per thousand. The inventory manager is evaluating such a policy: whenever the inventory level is equal to or below 2000, a order is pl4ced equal to the difference of the current inventory balance the maximum specified replenishment level of 4000.

Assuming that (b) the reorder point is (3000 + 70) units; (c) 20 weeks of data; (d) No back orders; (e) the orders are placed at the beginning of the week immediately after the inventory

level drops to (less than) the reorder point; (f) the replenishment orders are received at the beginning of the week.

Solution: Based on the weekly demand and lead time distributions, we will assign an appropriate set of random numbers to represent value or range of values of the variables given in tables 19.5 anti 19.6 respectively.

Table. 19.5

Demand/week (thousand)	Probability	Cumulative Probability	Random Number
0	0.2	0.2	00-19
1	0.4	0.6	20-59
2	0.3	0.9	60-89
3	0.1	1.0	90-99

Table 19.6

Lead Time (weeks)	Probability	Cumulative Probability	Random Number
2	0.3	0.3	00-29
3	0.4	0.7	30-69
4	0.3	1.0	70-99

Table 117 -The simulation of 20 weeks for the inventoryk system and related costs for the inventory system with a replenish level of 4000 units and the reorder level of 2000 units. The first random number is 31, this generates 1000 units of demand (from the cumulative probability values of calendar demand in table 115), leaving 2000 units on hand at the end of the first week. The last step shows it at the reorder level so an order of 4000-2000=2000 units is placed. Table 19.6 will help us get the lead time cumulative probability values, which gives us a value of 2 weeks at random n° 29. Assume 2000 units to be held, holding cos,k=Rs, 4 and the shortage cost is nil. In the succeeding week, the sequence 70 produces 2000 units of demand, corresponding to a possible value from table 19.5 and taking 2000 units on-hand at the start of the second week down to zero unit-:by the end of the week.

When in the third week the demand is for 1000 units, but as the available inventory is zero leads to the shortage cost of its. 10. Units ordered in first Week = 2000 and these are received in early fourth week. The fourth week demand is also 2000 units. and therefore always ending inventory is 0. So the second shortage happened in' sepet week and lasted till the end of eighth week. The orders placed at the end of week 4 are only received at the start of week 9. Thus, the simulation is carried on for 20 weeks}.

Summary: In order to evaluate the performance of the policy that has been simulated; we need to know: the number of orders placed, the average inventory, and the number of units short. As from table 19.7, during the course of simulation 5 times orders placed.

Table 19.7

REPLENISHMENT LEVEL = 4000 UNITS

Week	Beginning Inventory ('000)	Demand R.N.	Units	Ending Inventory ('000)	Lead Time R.N.	Lead Time Weeks	Quantity Ordered ('000)	Costs (Rs.) Holding	Costs (Rs.) Shortage
0				3	—	—	—	—	—
1	3	31	1	2	29	2	2	4	—
2	2	70	2	0	—	—	—	—	—
3	0	53	1	-1	—	—	—	—	10
4	2*	86	2	0	83	4	4	—	—
5	0	32	1	-1	—	—	—	—	10
6	0	78	2	-2	—	—	—	—	20
7	0	26	1	-1	—	—	—	—	10
8	0	64	2	-2	—	—	—	—	20
9	4*	45	1	3	—	—	—	—	—
10	3	12	0	3	—	—	—	6	—
11	3	99	3	0	58	3	4	6	—
12	0	52	1	-1	—	—	—	—	10
13	0	43	1	-1	—	—	—	—	10.
14	0	84	2	-2	—	—	—	—	20
15	4*	38	1	3	—	—	—	6	—
16	3	40	1	2	41	3	2	4	—
17	2	19	0	2	—	—	—	4	—
18	2	87	2	0	—	—	—	—	—
19	0	83	2	-2	—	—	—	—	20
20	2*	73	2	0	13	2	4	—	—

* Includes order quantity just received.

"The average inventory can be calculated by adding the weekly ending inventory balances (ignoring negative balances) and dividing by the number of weeks. Thus
Average inventory = 15000/20 = = 750 units per week
The total average weekly cost can be calculated as follows :
Weekly average cost = Ordering cost + Inventory holding cost + Shortage cost.
= (Rs. 50) (5) / 20 + (Rs. 2) (750)/ 1000 + (Rs. 10) (13) / 20
= 250 / 20 + 1.50 + 130.0/20
= Rs. 12.50 + 1.50 + 6.50
= Rs. 20.50
Minimum Stocking ----> In this case average shortage cost is much higher than holding cost. Thus, we can reduce this shortage cost by increasing the reorder level.

The two decision variable, replenishment level and reorder level, interact -with each other and influence the three. cost elements. Both, explanation of replenishment level and reorder level are interdependent, the experimentation done with the simulation model, should be actually performed in such a presentation so that various combinations of both variables can be recorded. Average lead time = - 2.8 weeks Average demand = 1400 units/week So average demand during lead time is 3920 units. Also a maximum lead time of 4 weeks and a maximum demand per week of 3000. So, during the lead time, the maximum demand equals 12000. Hence, the optimal reorder point, regardless of any replenishment level, must reside within the interval [3920 12000] units.

19.7 COMPUTER SIMULATION:

It is difficult, if not impossible, to perform •simulations without a computer. Consider the hospital simulation in this unit. Imagine what work would be involved if the analyst

simulated that one operating room for a month or simulated the entire 12- opemtkg rooms in the hospital for a months time Because hand-computed simulations are so expensive and so tedious, real simulations are done almost exclusively on a computer.

One of the most effective computer simulation languages is GPSS (General Purpose System Simulation) developed by IBM. One can also use computer language like FORTRAN to perform simulation experiments..

19.8 SUMMARY:

At the outset, a comprehensive overview of simulation has been provided through an introduction. We then moved on to the question of "why do you need simulation?" Here reasons for resorting to the technique of simulation have been given.

Every method has certain strengths and limitations. It is important for any one to understand the merits and shortcomings of a tool before actually using it. The limitations of simulation have also been covered

After providing the steps of simulation in a sequential manner, some typical applications of simulation have been illustrated which include-Home-Heating oil simulation, Carpet-cutting application, and Public School Planning system.

Then two typical examples have been manually solved using the standard method of simulation from the area of queuing and inventory in a step-by step manner. The idea here, is to provide a conceptual framework of solving a simulation model.

We have mentioned in a brief manner, the role of computers in simulation and emphasized the paramount importance of resorting to simulation solution through standard computer package like GPSS of IBM. It is impossible to do a complex simulation exercise without a computer in-practice.

19.9 TECHNICAL TERMS:

- **Model:** A representation of a system, process, or concept, often using mathematical or logical formulations.
- **Simulation:** The imitation of the operation of a real-world process or system over time.
- **System:** A set of interrelated components working together toward a common objective.
- **Entity:** The basic element being simulated (e.g., a person, a car, a machine).
- **Event**: A change in state that occurs at a point in time.
- **Random Variable**: A variable whose values are subject to variations due to chance.
- **Distribution**: Describes the probability of different outcomes (e.g., normal, exponential).
- **Validation:** Checking if the model accurately represents the real-world system.
- **Monte Carlo Simulation:** A method using random sampling to estimate statistical properties.
- **System Dynamics:** Uses stocks, flows, and feedback loops to model complex systems.

19.10 SELF-ASSESSMENT' PROBLEMS:

1. Separate solutions obtained from analytical models from those obtained from simulation models.
2. "The moment when it gets impossible to use an optimization technique for solving a problem, one has to rely on simulation technique. Discuss.
3. Explain the strengths and weaknesses of simulation clearly.
4. What are the steps to simulate?
5. Analysis in all scenarios, simulator without computer? Discuss.
6. The following frequency distribution was obtained for the number of unemployed individuals arriving at a one-person state unemployment office to collect their unemployment compensation cheque.

Inter-arrival Time (min)	Frequency	Service Time (min)	Frequency
2	10	2	10
3	20	3	20
4	40	4	40
5	20	5	20
6	10	6	10

The state office would like to simulate the operating characteristics of this one person state unemployment office during the normal operating hours of 10.00 a.m. to 11.00 a.m. Use the simulation to find the average wait time in the system, and the average time in the system, and the maximum queue length.

19.11 FURTHER READINGS:

1. Levin R. and. C.A. Kirkpatrick-Quantitative Approaches to Management, New York: Mcgraw Hill Book Company, 1975.
2. M.P. Gupta and J.K. Sharma Operations Research for Management, National Publishing House, New Delhi, 1987.

Dr Pachala Vijaya Vani

Appendix 1

Random Number Table (2500 Random Digits)

1581922396	2068577984	8262130892	8374856049	4637567488
0928105582	7295088379	9586111652	7055508767	6472382934
4112077556	3440672486	1882412963	0684012006	0933147914
7457477468	5435810788	9670852913	1291265730	4890031305
0099520858	3090908872	2039593181	5973470495	9776135501
7245174840	2275698645	8416549348	4676463101	2229367983
6749420382	4832630032	5670984959	5432114610	2966095680
5503161011	7413686599	1198757693	0414294470	0140121398
7164238934	7666127259	5263097712	5133648980	4011966963
3593969523	0272759769	0385998136	9999089966	7544056832
4192054466	0700014629	5169439659	8408705169	1074373131
9697426117	6488888550	4031652526	8123543276	0927534537
2007950579	9564268448	3457416988	1531027886	7016633739
4584768758	2389278610	3859431781	3643768456	4141314518
3840145867	9120831830	7228367652	1267173884	4020651657
0190453442	4800088084	1165628559	5407921254	3768932478
6766554338	5585265145	5089052204	9780623691	2195448096
6315116284	9172824179	5544814339	0016943666	3828538786
3908771938	4035554324	0840126299	4942039208	1475623997
5570024586	9324732596	1186563397	4425143199	3216653251
2999997185	0135965938	7678931194	1351031403	6002561840
7864375912	8383232768	1892857070	2323673751	3188881718
7065492027	6349104233	3382569662	4579426926	1513082455
0654683246	4765104877	8149224168	5468631609	6474393896
7830555058	5233147182	3519287786	2481675649	8907598697
7626984369	4725370390	9641916289	5049082870	7463807244
4785048453	3646121751	8436077768	2928794356	9956043516
4627791048	5765558107	8762592043	6183670830	6363845920
9376470693	0441608934	8749472723	2202271078	5897002653
1227991661	7936797034	9527342791	4711871173	8300978148
5582095589	5535798279	4764439855	6279247618	4446895088
4959397698	1056981450	8416606706	8234013222	6426813469
1824779358	1333750468	9434074212	5273692238	5902177065
7041092295	5726289716	3420847871	1820481234	0318831723
3555104281	0903099163	6827824899	6383872737	5901682626
9717595534	1634107293	8521057472	1471300754	3044151557
5571564123	7344613447	1129117244	3208461091	1699403490
4674262892	2809456764	5806554509	8224980942	3738031833
8461228715	0746980892	9285305274	6331989646	8764467686
1838538678	3049068967	6955157269	5482964330	2161984904
1834182305	6203476893	5937802079	3445230195	3694915658
1884227732	2923727501	8044389132	3511203081	6072112445
6791857341	6696243386	2219599137	3193884236	8224729718
3007929946	4031562749	5570757297	6273785046	1455349704
6085440621	2875556938	5496629750	4841817356	1443167141
7005051056	3496332071	5054070890	7303867953	6255181190
9846413446	8306646692	0661684251	8875127201	6251533454
0625457703	4229164694	7321363715	7051128285	1108468072
5457593922	9751489574	1799906380	1989141062	5595364247
4076486653	8950826528	4934582003	4071187742	1456207629

104IB26

ORIGINALITY REPORT

7%	5%	2%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Central University Of Andhra Pradesh Student Paper	7%
2	www.coursehero.com Internet Source	<1%
3	zombiedoc.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		