SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – I –PROBABILITY AND RANDOM PROCESSES– SMTA5205

Theory of Probability

Introduction

If an experiment is repeated under essential homogeneous and similar conditions we generally come across two types of situations:

(i) The result or what is usually known as the 'outcome' is unique or certain.

(ii) The result is not unique but may be one of the several possible outcomes.

The phenomena covered by (i) are known as deterministic. For example, for a perfect gas, PV = constant.

The phenomena covered by (ii) are known as probabilistic. For example, in tossing a coin we are not sure if a head or tail will be obtained.

In the study of statistics we are concerned basically with the presentation and interpretation of chance outcomes that occur in a planned study or scientific investigation.

Definition of various terms

Trial and event: Consider an experiment which, though repeated under essentially identical conditions, does not give unique results but may result in any one of the several possible outcomes. The experiment is known as a trial and outcomes are known as events or cases. For example, throwing of a die is a trial and getting 1(or 2 or … 6) is an event.

Exhaustive events: The total number of possible outcomes in any trial is known as exhaustive events or exhaustive cases. For example, in tossing of a coin there are two exhaustive case, viz.: Head and Tail(the possibility of the coin standing on an edge being ignored)

Favourable events or cases: The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event. For example, in throwing of two dice, the number of cases favourable to getting the sum 3 is: (1,2) and (2,1)

Mutually exclusive events: Events are said to be mutually exclusive or incompatible if the happening of any one of them precludes the happening of all the others, that is if no two or more of them can happen simultaneously in the same trial. For example, in tossing a coin the events head and tail are mutually exclusive.

Equally likely events: Outcomes of a trial are said to be equally likely, if taking into consideration all the relevant evidences, there is no reason to expect one in preference to the others. For example, in throwing an unbiased die, all the six faces are equally likely to come.

Sample Space: Consider an experiment whose outcome is not predictable with certainty. However, although the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known. This set of all possible outcomes of an experiment is known as the **sample space** of the experiment and is denoted by S.

Some examples follow.

1. If the outcome of an experiment consists in the determination of the sex of a newborn child, then

$$S = \{ g,b\}$$

where the outcome g means that the child is a girl and b that it is a boy.

2. If the experiment consists of flipping two coins, then the sample space consists of the following four points:

$$S = \{(H,H), (H,T), (T,H), (T,T)\}$$

The outcome will be (H,H) if both coins are heads, (H,T) if the first coin is heads and the second tails, (T,H) if the first is tails and the second heads, and (T,T) if both coins are tails.

3. If the experiment consists of tossing two dice, then the sample space consists if the 36 points

$$S = \{ (i,j):i,j = 1, 2, 3, 4, 5,$$

$$= \{ (1,1)------(1,6)-----(6,1)-----(6,6) \}$$

where the outcome (i,j) is said to occur if i appears on the leftmost die and j on the other die.

3.2. Definitions of Probability

1. Mathematical or Classical or a priori probability:

If a trial results in n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to the happening of an event E, then the probability 'p' of happening of E is given by,

$$p = P(E) = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of cases}} = \frac{m}{n}$$

2. Statistical or empirical probability:

If a trial is repeated a number of times under essentially homogenous and identical conditions, then the limiting value of the number of times the event happens to the number of trials, as the number of trials become indefinitely large is called the probability of happening of the event. Symbolically, if in n trials an event E happens m times, then the probability 'p' of the happening of E is given by,

$$P = P(E) = \lim_{n\to\infty} \frac{m}{n}$$

3. Axiomatic Definition:

Consider an experiment whose sample space is S. For each event E of the sample space S, we assume that a number P(E) is defined and satisfies the following three axioms.

Axiom 1: $0 \le P(E) \le 1$

Axiom 2: $P(S) = 1$

Axiom 3: For any sequence of mutually exclusive events, E_1, E_2, … (that is, events for which $E_i E_j = \Phi$, when $i \neq j$),

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Some Important Formulas

1. If A and B are any two events, then

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

This rule is known as additive rule on probability.

For three events A, B and C, we have,

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

2. If A and B are mutually exclusive events, then

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

In general, if A_1, A_2, … , A_n are mutually exclusive, then

$$P(A_1 \cup A_2 \cup A_3 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$$

3. If A and A^c are complementary events, then

$P(A) + P(A^c) = 1$

4. $P(S) = 1$

5. $P(\Phi) = 0$

6. If A and B are any two events, then

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

7. If A and B are independent events, then

$$P(A \cap B) = P(A) \times P(B)$$

Glossary of Probability terms:

Statement	Meaning in terms of Set theory
1. At least one of the events A or B occurs	$\omega \in A \cup B$
2. Both the events A and B occur	$\omega \in A \cap B$
3. Neither A nor B occurs	$\omega \in \overline{A} \cap \overline{B}$
4. Event A occurs and B does not occur	$\omega \in A \cap \overline{B}$
5. Exactly one of the events A or B occurs	$\omega \in A \Delta B$
6. If event A occurs, so does B	$A \subset B$
7. Events A and B are mutually exclusive	$A \cap B = \Phi$
8. Complementary event of A	$\overline{A}$

9. Sample space	Universal set S

Example 1: Find the probability of getting a head in tossing a coin.

Solution: When a coin is tossed, we have the sample space $\{$Head, Tail$\}$

Therefore, the total number of possible outcomes is 2

The favourable number of outcomes is 1, that is the head.

$\therefore$ The required probability is ½.

Example 2: Find the probability of getting two tails in two tosses of a coin.

Solution: When two coins are tossed, we have the sample space $\{$HH, HT, TH, TT$\}$

Where H represents the outcome Head and T represents the outcome Tail.

The total number of possible outcomes is 4.

The favourable number of outcomes is 1, that is TT

$\therefore$ The required probability is ¼.

Example 3: Find the probability of getting an even number when a die is thrown

Solution: When a die is thrown the sample space is $\{1, 2, 3, 4, 5, 6\}$

The total number of possible outcomes is 6

The favourable number of outcomes is 3, that is 2, 4 and 6

$\therefore$ The required probability is= $\frac{3}{6} = ½$.

Example 4: What is the chance that a leap year selected at random will contain 53 Sundays?

Solution: In a leap year(which consists of 366 days) there are 52 complete weeks and 2 days over. The following are the possible combinations for these two over days:

(i) Sunday and Monday (ii)Monday and Tuesday (iii)Tuesday and Wednesday (iv)Wednesday and Thursday (v)Thursday and Friday (vi)Friday and Saturday (vii)Saturday and Sunday.

In order that a leap year selected at random should contain 53 Sundays, one of the two over days must be Sunday. Since out of the above 7 possibilities, 2 viz. (i) and (ii)are favourable to this event,

Required probability $= \frac{2}{7}$

Example 5: If two dice are rolled, what is the probability that the sum of the upturned faces will equal 7?

Solution: We shall solve this problem under the assumption that all of the 36 possible outcomes are equally likely. Since there are 6 possible outcomes – namely (1,6), (2,5), (3,4), (4,3), (5,2,), (6,1) – that result in the sum of the dice being equal to 7, the desired probability is $\frac{6}{36} = \frac{1}{6}$.

Example 6: A bag contains 3 Red, 6 White and 7 Blue balls. What is the probability that two balls drawn are white and blue?

Solution: Total number of balls = 3 + 6 + 7 = 16.

Out of 16 balls, 2 can be drawn in $16C_2$ ways.

Therefore exhaustive number of cases is 120.

Out of 6 white balls 1 ball can be drawn in $6C_1$ ways and out of 7 blue balls 1 ball can be drawn in $7C_1$ ways. Since each of the former cases can be associated with each of the latter cases, total number of favourable cases is $6C_1 \times 7C_1 = 6 \times 7 = 42$.

$\therefore$ The required probability is = $\frac{42}{120} = \frac{7}{20}$

Example 7: A lot consists of 10 good articles, 4 with minor defects and 2 with major defects. Two articles are chosen from the lot at random (without replacement). Find the probability that (i) both are good, (ii) both have major defects, (iii) at least 1 is good, (iv) at most 1 is good, (v)exactly 1 is good, (vi) neither has major defects and (vii) neither is good.

Solution: Although the articles may be drawn one after the other, we can consider that both articles are drawn simultaneously, as they are drawn without replacement.

(i) $P(both \text{ are good}) = \frac{\text{No. of ways drawing 2 good articles}}{\text{Total no. of ways of drawing 2 articles}}$

$= \frac{10C_2}{16C_2} = \frac{3}{8}$

(ii) $P(\text{both have major defects}) = \dfrac{\text{No. of ways of drawing 2 articles with major defects}}{\text{Total no. of ways}}$

$$= \frac{2C_2}{16C_2} = \frac{1}{120}$$

(iii) P(at least 1 is good) = P(exactly 1 is good or both are good)

=P(exactly 1 is good and 1 is bad or both are good)

$$= \frac{10C_1 x 6C_1 + 10C_2}{16C_2} = \frac{7}{8}$$

(iv) P(atmost 1 is good) =P(none is good or 1 is good and 1is bad)

$$= \frac{10C_0 x 6C_2 + 10C_1 x 6C_1}{16C_2} = \frac{5}{8}$$

(v) P(exactly 1is good) =P(1 is good and 1 is bad)

$$= \frac{10C_1 x 6C_1}{16C_2} = \frac{1}{2}$$

(vi) P(neither has major defects) = P(both are non-major defective articles)

$$= \frac{14C_2}{16C_2} = \frac{91}{120}$$

(vii) P(neither is good) = P(both are defective)

$$= \frac{6C_2}{16C_2} = \frac{1}{8}$$

Example 8: From 6 positive and 8 negative numbers, 4 numbers are chosen at random (without replacement) and multiplied. What is the probability that the product is positive?

Solution: If the product is to be positive, all the 4 numbers must be positive or all the 4 must be negative or 2 of them must be positive and the other 2 must be negative.

No. of ways of choosing 4 positive numbers= $6C_4$ =15.

No. of ways of choosing 4 negative numbers= $8C_4$ =70.

No.of ways of choosing 2 positive and 2 negative numbers

$= {}^{6}C_{2}\, x\, {}^{8}C_{2} = 420.$

Total no. of ways of choosing 4 numbers from all the 14 numbers

$= {}^{14}C_{4} = 1001.$

P(the product is positive)

$$= \frac{No.\text{ of ways by which t he product is positive}}{\text{Total no. of ways}}$$

$$= \frac{15 + 70 + 420}{1001} = \frac{505}{1001}$$

Example 9: If 3 balls are "randomly drawn" from a bowl containing 6 white and 5 black balls, what is the probability that one of the drawn balls is white and the other two black?

Solution: If we regard the order in which the balls are selected as being relevant, then the sample space consists of 11· 10 · 9 = 990 outcomes. Furthermore, there are 6· 5· 4 = 120 outcomes in which the first ball selected is white and the other two black; 5 · 6· 4 = 120 outcomes in which the first is black, the second white and the third black; and 5· 4 · 6 = 120 in which the first two are black and the third white. Hence, assuming that "randomly drawn" means that each outcome in the sample space is equally likely to occur, we see that the desired probability is $\frac{120+120+120}{990} = \frac{4}{11}$

Example 10: In a large genetics study utilizing guinea pigs, *Cavia sp.,* 30% of the offspring produced had white fur and 40% had pink eyes. Two-thirds of the guinea pigs with white fur had pink eyes. What is the probability of a randomly selected offspring having both white fur and pink eyes?

Solution: P(W) = 0.30, P(Pi) = 0.40, and P(Pi W) = 0.67. Utilizing Formula 2.9,

P(Pi ∩ W) = P(Pi W). P(W) = 0.67. 0.30 = 0.20.

Twenty percent of all offspring are expected to have both white fur and pink eyes.

Example 11: Consider three gene loci in tomato, the first locus affects fruit shape with the oo genopyte causing oblate or flattened fruit and OO or Oo normal round fruit. The second locus affects fruit color with yy having yellow fruit and YY or Yy red fruit. The final locus affects leaf shape with pp having potato or smooth leaves and PP or Pp having the more typical cut leaves. Each of these loci is located on a different pair of chromosomes and, therefore, acts independently of the other loci. In the following cross OoYyPp × OoYypp, what is the probability that an offspring will have the dominant phenotype for each trait? What is the probability that it will be heterozygous for all three genes? What is the probability that it will have round, yellow fruit and potato leaves?

Solution: Genotypic array:

$$(\frac{1}{4}\text{OO} + \frac{2}{4}\text{Oo} + \frac{1}{4}\text{oo})\,(\frac{1}{4}\text{YY} + \frac{2}{4}\text{Yy} + \frac{1}{4}\text{yy})\,(\frac{1}{2}\text{pp})$$

Phenotypic array:

$$(\frac{3}{4}\text{O-} + \frac{1}{4}\text{oo})\,(\frac{3}{4}\text{Y-} + \frac{1}{4}\text{yy})\,(\frac{1}{2}\text{P} + \frac{1}{2}\text{pp})$$

The probabiltity of dominant phenotype for each trait from the phenotypic array above is

$$\text{P(O-Y-P-)} = \text{P(O-)} \times \text{P(Y-)} \times \text{P(P-)} = \frac{3}{4} \times \frac{3}{4} \times \frac{1}{2} = \frac{9}{32}.$$

The probability of heterozygous for all three genes from the genotypic array above is

$$\text{P(OoYyPp)} = \text{P(Oo)} \times \text{P(Yy)} \times \text{P(Pp)} = \frac{2}{4} \times \frac{2}{4} \times \frac{1}{2} = \frac{4}{32} = \frac{1}{8}.$$

The probability of a round, yellow-fruited plant with potato leaves from the phenotypic array above is

$$\text{P(O-yypp)} = \text{P(O-)} \times \text{P(yy)} \times \text{P(pp)} = \frac{3}{4} \times \frac{1}{4} \times \frac{1}{2} = \frac{3}{32}.$$

Each answer applies the probability rules for independent events to the separate gene loci.

Example 12: (a) Two cards are drawn at random from a well shuffled pack of 52 playing cards. Find the chance of drawing two aces.

(b) From a pack of 52 cards, three are drawn at random. Find the chance that they are a king, a queen and a knave.

(c) Four cards are drawn from a pack of cards. Find the probability that (i) all are diamond (ii) there is one card of each suit (iii) there are two spades and two hearts.

Solution: (a) From a pack of 52 cards 2 can be drawn in $52C_2$ ways, all being equally likely. $\therefore$ Exhaustive number of cases is $52C_2$.

In a pack there are 4 aces and therefore 2 aces can be drawn in $4C_2$ ways.

$\therefore$ Required probability = $\frac{4C_2}{52C_2} = \frac{1}{221}$

(b) Exhaustive number of cases = $52C_3$

A pack of cards contains 4 kings, 4 queens and 4 knaves. A king, a queen and a knave can each be drawn in $4C_1$ ways and since each way of drawing a king can be associated with each of the ways of drawing a queen and a knave, the total number of favrourable cases = $4C_1 \times 4C_1 \times 4C_1$.

$\therefore$ Required probability = $\frac{4C_1 \times 4C_1 \times 4C_1}{52C_3} = \frac{16}{5525}$

(c) Exhaustive number of cases $52C_4$

(i) Required probability = $\frac{13C_4}{52C_4}$

(ii) Required probability = $\frac{13C_1 \times 13C_1 \times 13C_1 \times 13C_1}{52C_4}$

(iv) Required probability = $\frac{13C_2 \times 13C_2}{52C_4}$

Example 13: What is the probability of getting 9 cards of the same suit in one hand at a game of bridge?

Solution: One hand in a game of bridge consists of 13 cards.

$\therefore$ Exhaustive number of cases $52C_{13}$

Number of ways in which, in one hand, a particular player gets 9 cards of one suit are $13C_9$ and the number of ways in which the remaining 4 cards are of some other suit are $39C_4$. Since there are 4 suits in a pack of cards, total number of favourable cases is $4 \times 13C_9 \times 39C_4$.

$\therefore$ Required probability = $\frac{4 \times 13C_9 \times 39C_4}{52C_{13}}$

Example 14: A committee of 4 people is to be appointed from 3 officers of the production department, 4 officers of the purchase department, two officers of the sales department and 1 chartered accountant. Find the probability of forming the committee in the following manner:

(i) There must be one from each category
(ii) It should have at least one from the purchase department
(iii) The chartered accountant must be in the committee.

Solution: There are 3 + 4 + 2 + 1 = 10 persons in all and a committee of 4 people can be formed out of them in $10C_4$ ways. Hence exhaustive number of cases is $10C_4$ = 210

(i) Favourable number of cases for the committee to consist of 4 members, one from each category is $4C_1 \times 3C_1 \times 2C_1 \times 1$ = 24

$\therefore$ Required probability = $\frac{24}{120}$

(ii) P(Committee has at least one purchase officer) = 1 – P(Committee has no purchase Officer)

In order that the committee has no purchase officer, all the four members are to be selected amongst officers of production department, sales department and chartered accountant, that is out of 3 + 2 + 1 = 6 members and this can be done in $^{5}C_{4}$ = 15 ways. Hence,

P(Committee has no purchase officer) = $\frac{15}{210} = \frac{1}{14}$

$\therefore$ P(Committee has at least one purchase officer) = 1 – $\frac{1}{14}$ = $\frac{13}{14}$

(iii) Favourable number of cases that the committee consists of a chartered accountant as a member and three others are:

$1 \times {}^{9}C_{3}$ = 84 ways.

Since a chartered accountant can be selected out of one chartered accountant in only 1 way and the remaining 3 members can be selected out of the remaining 10 – 1 persons in $^{9}C_{3}$ ways. Hence the required probability = $\frac{84}{210} = \frac{2}{5}$.

Example 15: A box contains 6 red, 4 white and 5 black balls. A persons draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each colour.

Solution: The required event E that in a draw of 4 balls from the box at random there is at least one ball of each colour can materialize in the following mutually disjoint ways:

(i) 1 Red, 1 White and 2 Black balls
(ii) 2 Red, 1 White and 1 Black balls
(iii) 1 Red, 2 White and 1 Black balls

Hence by addition rule of probability, the required probability is given by,

P(E) = P(i) + P(ii) + P(iii)

$$= \frac{6C_1 \times 4C_1 \times 5C_2}{15C_4} + \frac{6C_2 \times 4C_1 \times 5C_1}{15C_4} + \frac{6C_1 \times 4C_2 \times 5C_1}{15C_4}$$

= 0.5275

Example 16: A problem in Statistics is given to the three students A, B and C whose chances of solving it are 1/2, 3/4 and 1/4 respectively. What is the probability that the problem will be solved if all of them try independently?

Solution: Let A, B and C denote the events that the problem is solved by the students A, B and C respectively. Then

P(A) = 1/2 P(B) = 3/4 P(C) = 1/4

P($\overline{A}$) = 1 – 1/2 = 1/2 P($\overline{B}$) = 1 – 3/4 = 1/4 P($\overline{C}$) = 1 – 1/4 = 3/4

P(Problem solved) = P(At least one of them solves the problem)

= 1 – P(None of them solve the problem)

= 1 – P($\overline{A \cup B \cup C}$)

= 1 – P($\overline{A} \cap \overline{B} \cap \overline{C}$)

= 1 – P($\overline{A}$) P($\overline{B}$) P($\overline{C}$)

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

$= \frac{29}{32}$

Example 17: Three groups of children contain respectively 3 girls and 1 boy, 2 girls and 2 boys and 1 girl and 3 boys. One child is selected at random from each group. Find the probability that the three selected consist of 1 girl and 2 boys.

Solution: The required event of getting 1 girl and 2 boys among the three selected children can materialize in the following three mutually exclusive cases:

Group No. →	**I**	**II**	**III**
(i)	**Girl**	**Boy**	**Boy**
(ii)	**Boy**	**Girl**	**Boy**
(iii)	**Boy**	**Boy**	**Girl**

By addition rule of probability,

Required probability = P(i) + P(ii) + P(iii)

Since the probability of selecting a girl from the first group is 3/4, of selecting a boy from the second is 2/4, and of selecting a boy from the third group is ¾, and since these three events of selecting children from the three groups are independent of each other, we have,

P(i) = $\frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{9}{32}$

P(ii) = $\frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{3}{32}$

P(iii) = $\frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{1}{32}$

Hence the required probability = $\frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}$

Conditional Probability and Baye's Theorem

Conditional Probability and Multiplication Law

For two events A and B

$P(A \cap B) = P(A) \cdot P(B/A), P(A) > 0$

$= P(B) \cdot P(A/B), P(B) > 0$

where P(B/A) represents the conditional probability of occurrence of B when the event A has already happened and P(A/B) is the conditional probability of occurrence of A when the event B has already happened.

Theorem of Total Probability:

If $B_1, B_2, \ldots, B_n$ be a set of exhaustive and mutually exclusive events, and A is another event associated with (or caused by) B_i, then

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A/B_i)$$

Example 18 : A box contains 4 bad and 6 good tubes. Two are drawn out from the box at a time. One of them is tested and found to be good. What is the probability that the other one is also good?

Solution: Let A = one of the tubes drawn is good and B = the other tube is good.

P(A∩B) = P(both tubes drawn are good)

$$= \frac{6C_2}{10C_2} = \frac{1}{3}$$

Knowing that one tube is good, the conditional probability that the other tube is also good is required, i.e., P(B/A) is required.

By definition,

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{6/10} = \frac{5}{9}$$

Example 19: A bolt is manufactured by 3 machines A, B and C. A turns out twice as many items as B, and machines B and C produce equal number of items. 2% of bolts produced by A and B are defective and 4% of bolts produced by C are defective. All bolts are put into 1 stock pile and chosen from this pile. What is the probability that it is defective?

Solution: Let A = the event in which the item has been produced by machine A, and so on.

Let D = the event of the item being defective.

$P(A) = \frac{1}{2}$, $P(B) = P(C) = \frac{1}{4}$

P(D/A) = P(an item is defective, given that A has produced it)

$= \frac{2}{100}$ = P(D/B)

$P(D/C) = \frac{4}{100}$

By theorem of total probability,

P(D) = P(A)× P(D/A) + P(B) × P(D/B) + P(C) ×P(D/c)

$= \frac{1}{2} \times \frac{2}{100} + \frac{1}{4} \times \frac{2}{100} + \frac{1}{4} \times \frac{4}{100}$

$= \frac{1}{40}$

Example 20: In a coin tossing experiment, if the coin shows head, one die is thrown and the result is recorded. But if the coin shows tail, 2 dice are thrown and their sum is recorded. What is the probability that the recorded number will be 2?

Solution: When a single die is thrown, P(2) = 1/6

When 2 dice are thrown, the sum will be 2 only if each dice shows 1.

$\therefore$ P(getting 2 as sum with 2 dice) = $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$ (since independence)

By theorem of total probability,

$$P(2) = P(H) \times P(2/H) + P(T) \times P(2/T)$$

$$= \frac{1}{2}\times\frac{1}{6}+\frac{1}{2}\times\frac{1}{36}=\frac{7}{72}$$

Example 21: An urn contains 10 white and 3 black balls. Another urn contains 3 white and 5 black balls. Two balls are drawn at random from the first urn and place in the second urn and then one ball is taken at random from the latter. What is the probability that it is a white ball?

Solution: The two balls transferred may be both white or both black or one white and one black.

Let B_1 = event of drawing 2 white balls from the first urn, B_2 = event of drawing 2 black balls from it and B_3 = event of drawing one white and one black ball from it.

Clearly B_1, B_2 and B_3 are exhaustive and mutually exclusive events.

Let A = event of drawing a white ball from the second urn after transfer.

$$P(B_1) = \frac{10C_2}{13C_2}=\frac{15}{26}$$

$$P(B_2) = \frac{3C_2}{13C_2}=\frac{1}{26}$$

$$P(B_3) = \frac{10\times 3}{13C_2}=\frac{10}{26}$$

$P(A/B_1)$ = P(drawing a white ball / 2 white balls have been transferred)

= P(drawing a white ball / urn II contains 5 white and 5 black balls)

$$= \frac{5}{10}$$

Similarly, $P(A/B_2) = \frac{3}{10}$ and $P(A/B_3) = \frac{4}{10}$

By theorem of total probability,

$$P(A) = P(B_1) \times P(A/B_1) + P(B_2) \times P(A/B_2) + P(B_3) \times P(A/B_3)$$

$$= \frac{15}{26}\times\frac{5}{10}+\frac{1}{26}\times\frac{3}{10}+\frac{10}{26}\times\frac{4}{10}=\frac{59}{130}$$

Example 22: In 1989 there were three candidates for the position of principal – Mr.Chatterji, Mr. Ayangar and Mr. Singh – whose chances of getting the appointment are in the proportion 4:2:3 respectively. The probability that Mr. Chatterji if selected would introduce co-education in the college is 0.3. The probabilities of Mr. Ayangar and Mr.Singh doing the same are respectively 0.5 and 0.8. What is the proabability that there will be co-education in the college?

Solution: Let the events and probabilities be defined as follows:

A: Introduction of co-education

E_1: Mr.Chatterji is selected as principal

E_2: Mr.Ayangar is selected as principal

E_3: Mr.Singh is selected as principal

Then,

$P(E_1) = \frac{4}{9}$ $P(E_2) = \frac{2}{9}$ $P(E_3) = \frac{3}{9}$

$P(A/E_1) = 0.3$ $P(A/E_2) = 0.5$ $P(A/E_3) = 0.8$

$P(A) = P\left[(A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3)\right]$

$= P\left[(A \cap E_1) + (A \cap E_2) + (A \cap E_3)\right]$

$= P(E_1)\, P(A/E_1) + P(E_2)\, P(A/E_2) + P(E_3)\, P(A/E_3)$

$= \frac{4}{9} \times \frac{3}{10} + \frac{2}{9} \times \frac{5}{10} + \frac{3}{9} \times \frac{8}{10} = \frac{23}{45}$

3.3.4. Baye's theorem

If $E_1, E_2, \ldots, E_n$ are mutually disjoint events with $P(E_i) \neq 0$, (i = 1,2, … , 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,

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

3.3.5. Solved Examples

Example 23. A bag contains 5 balls and it is not known how many of them are white. Two balls are drawn at random from the bag and they are noted to be white. What is the chance that all the balls in the bag are white?

Solution: Since 2 white balls have been drawn out, the bag must have contained 2, 3, 4 or 5 white balls.

Let B_1 = Event of the bag containing 2 white balls, B_2 = Events of the bag containing 3 white balls, B_3 = Event of the bag containing 4 white balls and B_4 = Event of the bag containing 5 white balls.

Let A = Event of drawing 2 white balls.

$P(A/B_1) = \frac{2C_2}{5C_2} = \frac{1}{10}$ $P(A/B_2) = \frac{3C_2}{5C_2} = \frac{3}{10}$

$P(A/B_3) = \frac{4C_2}{5C_2} = \frac{4}{10}$ $P(A/B_4) = \frac{5C_2}{5C_2} = 1$

Since the number of white balls in the bag is not known, B_i's are equally likely.

$P(B_1) = P(B_2) = P(B_3) = P(B_4) = \frac{1}{4}$

By Baye's theorem,

$$P(B_4/A) = \frac{P(B_4) \times P(A/B_4)}{\sum_{i=1}^{4} P(B_i) \times P(A/B_i)}$$

$$= \frac{\frac{1}{4} \times 1}{\frac{1}{4} \times \left(\frac{1}{10} + \frac{3}{10} + \frac{3}{5} + 1 \right)} = \frac{1}{2}$$

Example 24: There are 3 true coins and 1 false coin with 'head' on both sides. A coin is chosen at random and tossed 4 times. If 'head' occurs all the 4 times, what is the probability that the false coin has beeb chosen and used?

Solution:

$P(T) = P(\text{the coin is a true coin}) = \frac{3}{4}$

$P(F) = P(\text{the coin is a false coin}) = \frac{1}{4}$

Let A = Event of getting all heads in 4 tosses

Then P(A/T) = $\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$ and P(A/F) = 1

By Baye's theorem

P(F/A) = $\frac{P(F)\times P(A/F)}{P(F)\times P(A/F)+P(T)\times P(A/T)}$

= $\frac{\frac{1}{4}\times 1}{\frac{1}{4}\times 1+\frac{3}{4}\times\frac{1}{16}}=\frac{16}{19}$

Example 25: The contents of urns I, Ii and III are as follows:

1 white, 2 black and 3 red balls

2 white, 1 black and 1 red balls

4 white, 5 black and 3 red balls

One urn is chosen at random and two balls are drawn. They happen to be white and red. What is the probability that they come from urns I, II or III?

Solution: Let E_1, E_2 and E_3 denote the events that the urn I, II and III is chosen, respectively, and let A be the event that the two balls taken from the selected urn are white and red. Then

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

$P(A/E_1) = \frac{1\times 3}{6C_2}=\frac{1}{5}$

$P(A/E_2) = \frac{2\times 1}{4C_2}=\frac{1}{3}$

$P(A/E_3) = \frac{4\times 3}{12C_2}=\frac{2}{11}$

Hence $P(E_2/A) = \dfrac{P(E_2)P(A/E_2)}{\sum_{i=1}^{3} P(E_i)P(A/E_i)}$

$$= \frac{\frac{1}{3}\times\frac{1}{3}}{\frac{1}{3}\times\frac{1}{5}+\frac{1}{3}\times\frac{1}{3}+\frac{1}{3}\times\frac{2}{11}} = \frac{55}{118}$$

Similarly, $P(E_3/A) = \dfrac{\frac{1}{3}\times\frac{2}{11}}{\frac{1}{3}\times\frac{1}{5}+\frac{1}{3}\times\frac{1}{3}+\frac{1}{3}\times\frac{1}{11}} = \frac{30}{118}$

Therefore $P(E_1/A) = 1-\frac{55}{118}-\frac{30}{118} = \frac{33}{118}$

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

1. DISCRETE RANDOM VARIABLES

1.1. **Definition of a Discrete Random Variable.** A random variable X is said to be *discrete* if it can assume only a finite or countable infinite number of distinct values. A discrete random variable can be defined on both a countable or uncountable sample space.

1.2. **Probability for a discrete random variable.** The probability that X takes on the value x, P(X=x), is defined as the sum of the probabilities of all sample points in Ω that are assigned the value x. We may denote P(X=x) by p(x). The expression p(x) is a function that assigns probabilities to each possible value x; thus it is often called the probability function for X.

1.3. **Probability distribution for a discrete random variable.** The probability distribution for a discrete random variable X can be represented by a formula, a table, or a graph, which provides p(x) = P(X=x) for all x. The probability distribution for a discrete random variable assigns nonzero probabilities to only a countable number of distinct x values. Any value x not explicitly assigned a positive probability is understood to be such that P(X=x) = 0.

The function f(x) p(x)= P(X=x) for each x within the range of X is called the *probability distribution* of X. It is often called the probability mass function for the discrete random variable X.

1.4. **Properties of the probability distribution for a discrete random variable.** A function can serve as the probability distribution for a discrete random variable X if and only if it s values, f(x), satisfy the conditions:

a: f(x) $\geq$ 0 for each value within its domain
b: $\sum_x \ f(x) \ = \ 1$, where the summation extends over all the values within its domain

1.5. **Examples of probability mass functions.**

1.5.1. *Example 1.* Find a formula for the probability distribution of the total number of heads obtained in four tosses of a balanced coin.

The sample space, probabilities and the value of the random variable are given in table 1.
From the table we can determine the probabilities as

$$P(X=0) \ = \frac{1}{16} \, , \ P(X=1) \ = \frac{4}{16} \, , \ P(X=2) \ = \frac{6}{16} \, , \ P(X=3) \ = \frac{4}{16} \, , \ P(X=4) \ = \frac{1}{16} \qquad (1)$$

Notice that the denominators of the five fractions are the same and the numerators of the five fractions are 1, 4, 6, 4, 1. The numbers in the numerators is a set of binomial coefficients.

$$\frac{1}{16} \ = \ \binom{4}{0}\frac{4}{16} \ = \ \binom{4}{1}\frac{6}{16} \ = \ \binom{4}{2}\frac{4}{16} \ = \ \binom{4}{3}\frac{1}{16} \ = \ \binom{4}{4}$$

We can then write the probability mass function as

Date: August 20, 2004.

TABLE 1. **Probability of a Function of the Number of Heads from Tossing a Coin Four Times.**

Table R.1 Tossing a Coin Four Times		
Element of sample space	Probability	Value of random variable X (x)
HHHH	1/16	4
HHHT	1/16	3
HHTH	1/16	3
HTHH	1/16	3
THHH	1/16	3
HHTT	1/16	2
HTHT	1/16	2
HTTH	1/16	2
THHT	1/16	2
THTH	1/16	2
TTHH	1/16	2
HTTT	1/16	1
THTT	1/16	1
TTHT	1/16	1
TTTH	1/16	1
TTTT	1/16	0

$$f(x) = \frac{\binom{4}{x}}{16} \; for \; x = 0, 1, 2, 3, 4 \tag{2}$$

Note that all the probabilities are positive and that they sum to one.

1.5.2. *Example 2.* Roll a red die and a green die. Let the random variable be the larger of the two numbers if they are different and the common value if they are the same. There are 36 points in the sample space. In table 2 the outcomes are listed along with the value of the random variable associated with each outcome.

The probability that X = 1, P(X=1) = P[(1, 1)] = 1/36. The probability that X = 2, P(X=2) = P[(1, 2), (2,1), (2, 2)] = 3/36. Continuing we obtain

$$P(X=1) = \binom{1}{36}, \; P(X=2) = \binom{3}{36} \; P(X=3) = \binom{5}{36}$$
$$P(X=4) = \binom{7}{36}, \; P(X=5) = \binom{9}{36}, \; P(X=6) = \binom{11}{36}$$

We can then write the probability mass function as

$$f(x) = P(X=x) = \frac{2x-1}{36} \; for \; x = 1, 2, 3, 4, 5, 6$$

Note that all the probabilities are positive and that they sum to one.

1.6. **Cumulative Distribution Functions.**

TABLE 2. Possible Outcomes of Rolling a Red Die and a Green Die – First Number in Pair is Number on Red Die

Green (A)	1	2	3	4	5	6
Red (D)						
1	1 1 *1*	1 2 *2*	1 3 *3*	1 4 *4*	1 5 *5*	1 6 *6*
2	2 1 *2*	2 2 *2*	2 3 *3*	2 4 *4*	2 5 *5*	2 6 *6*
3	3 1 *3*	3 2 *3*	3 3 *3*	3 4 *4*	3 5 *5*	3 6 6
4	4 1 *4*	4 2 *4*	4 3 *4*	4 4 *4*	4 5 *5*	4 6 *6*
5	5 1 *5*	5 2 *5*	5 3 *5*	5 4 *5*	5 5 *5*	5 6 *6*
6	6 1 *6*	6 2 *6*	6 3 *6*	6 4 *6*	6 5 *6*	6 6 *6*

1.6.1. *Definition of a Cumulative Distribution Function.* If X is a discrete random variable, the function given by

$$F(x) = P(x \leq X) = \sum_{t \leq x} f(t) \; for \; -\infty \leq x \leq \infty \tag{3}$$

where f(t) is the value of the probability distribution of X at t, is called the *cumulative distribution function* of X. The function F(x) is also called the *distribution function* of X.

1.6.2. *Properties of a Cumulative Distribution Function.* The values F(X) of the distribution function of a discrete random variable X satisfy the conditions

1: F(-∞) = 0 and F(∞) =1;
2: If a $<$ b, then F(a) $\leq$ F(b) for any real numbers a and b

1.6.3. *First example of a cumulative distribution function.* Consider tossing a coin four times. The possible outcomes are contained in table 1 and the values of f in equation 1. From this we can determine the cumulative distribution function as follows.

$$\begin{aligned}
F(0) &= f(0) = \frac{1}{16} \\
F(1) &= f(0) + f(1) = \frac{1}{16} + \frac{4}{16} = \frac{5}{16} \\
F(2) &= f(0) + f(1) + f(2) = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} = \frac{11}{16} \\
F(3) &= f(0) + f(1) + f(2) + f(3) = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{6} = \frac{15}{16} \\
F(4) &= f(0) + f(1) + f(2) + f(3) + f(4) = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{6} + \frac{1}{16} = \frac{16}{16}
\end{aligned}$$

We can write this in an alternative fashion as

$$F(x) = \begin{cases} 0 & for\ x < 0 \\ \frac{1}{16} & for\ 0 \le x < 1 \\ \frac{5}{16} & for\ 1 \le x < 2 \\ \frac{11}{16} & for\ 2 \le x < 3 \\ \frac{15}{16} & for\ 3 \le x < 4 \\ 1 & for\ x \ge 4 \end{cases}$$

1.6.4. *Second example of a cumulative distribution function.* Consider a group of N individuals, M of whom are female. Then N-M are male. Now pick n individuals from this population without replacement. Let x be the number of females chosen. There are $\binom{M}{x}$ ways of choosing x females from the M in the population and $\binom{N-M}{n-x}$ ways of choosing n-x of the N - M males. Therefore, there are $\binom{M}{x} \times \binom{N-M}{n-x}$ ways of choosing x females and n-x males. Because there are $\binom{N}{n}$ ways of choosing n of the N elements in the set, and because we will assume that they all are equally likely the probability of x females in a sample of size n is given by

$$f(x) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} for\ x = 0,\ 1,\ 2,\ 3, \cdots,\ n \quad and\ x \le M,\ and\ n - x \le N - M. \tag{4}$$

For this discrete distribution we compute the cumulative density by adding up the appropriate terms of the probability mass function.

$$\begin{aligned} F(0) &= f(0) \\ F(1) &= f(0) + f(1) \\ F(2) &= f(0) + f(1) + f(2) \\ F(3) &= f(0) + f(1) + f(2) + f(3) \\ &\vdots \\ F(n) &= f(0) + f(1) + f(2) + f(3) + \cdots + f(n) \end{aligned} \tag{5}$$

Consider a population with four individuals, three of whom are female, denoted respectively by A, B, C, D where A is a male and the others are females. Then consider drawing two from this population. Based on equation 4 there should be $\binom{4}{2}$ = 6 elements in the sample space. The sample space is given by

TABLE 3. **Drawing Two Individuals from a Population of Four where Order Does Not Matter (no replacement)**

Element of sample space	Probability	Value of random variable X
AB	1/6	1
AC	1/6	1
AD	1/6	1
BC	1/6	2
BD	1/6	2
CD	1/6	2

We can see that the probability of 2 females is $\frac{1}{2}$. We can also obtain this using the formula as follows.

$$f(2) \;=\; P(X=\,2) \;=\; \frac{\binom{3}{2}\,\binom{1}{0}}{\binom{4}{2}} = \frac{(3)(1)}{6} = \frac{1}{2} \tag{6}$$

Similarly

$$f(1) \;=\; P(X=\,1) \;=\; \frac{\binom{3}{1}\,\binom{1}{1}}{\binom{4}{2}} = \frac{(3)(1)}{6} = \frac{1}{2} \tag{7}$$

We cannot use the formula to compute f(0) because (2 - 0) $\not\leq$ (4 - 3). f(0) is then equal to 0. We can then compute the cumulative distribution function as

$$\begin{aligned} F(0) \;&= f(0) \;=\; 0 \\ F(1) \;&= f(0) \;+\; f(1) \;= \frac{1}{2} \\ F(2) \;&= f(0) \;+\; f(1) \;+\; f(2) \;=\; 1 \end{aligned} \tag{8}$$

1.7. **Expected value.**

1.7.1. *Definition of expected value.* Let X be a discrete random variable with probability function p(x). Then the *expected value* of X, E(X), is defined to be

$$E(X) \;=\; \sum_{x} \; x\; p(x) \tag{9}$$

if it exists. The expected value exists if

$$\sum_{x} \; |\, x \,|\; p(x) \;<\; \infty \tag{10}$$

The expected value is kind of a weighted average. It is also sometimes referred to as the population mean of the random variable and denoted μ_X.

1.7.2. *First example computing an expected value.* Toss a die that has six sides. Observe the number that comes up. The probability mass or frequency function is given by

$$p\,(x) \;=\; P(X \;=\; x) \;= \begin{cases} \frac{1}{6} & for\; x \;=\; 1,\, 2,\, 3,\, 4,\, 5,\, 6 \\ 0 & otherwise \end{cases} \tag{11}$$

We compute the expected value as

$$\begin{aligned} E(X) \;&= \sum_{x \;\epsilon\; X} \; x\; p_X\,(x) \\ &= \sum_{i\;=\;1}^{6} \; i \;\left(\frac{1}{6}\right) \\ &= \frac{1 + \;2\; +\; 3\; +\; 4\; +\; 5\; +\; 6}{6} \\ &= \frac{21}{6} \;=\; 3\;\frac{1}{2} \end{aligned} \tag{12}$$

1.7.3. *Second example computing an expected value.* Consider a group of 12 television sets, two of which have white cords and ten which have black cords. Suppose three of them are chosen at random and shipped to a care center. What are the probabilities that zero, one, or two of the sets with white cords are shipped? What is the expected number with white cords that will be shipped?

It is clear that x of the two sets with white cords and 3-x of the ten sets with black cords can be chosen in $\binom{2}{x} \times \binom{10}{3-x}$ ways. The three sets can be chosen in $\binom{12}{3}$ ways. So we have a probability mass function as follows.

$$f(x) \;=\; P(\,X \;=\; x) \;=\; \frac{\binom{2}{x}\,\binom{10}{3-x}}{\binom{12}{3}} \;\; for\; x \;=\; 0\,,\; 1\,,\; 2 \tag{13}$$

For example

$$f(x) \;=\; P(X \;=\; x) \;=\; \frac{\binom{2}{0}\,\binom{10}{3-0}}{\binom{12}{3}} \;=\; \frac{(1)\,(120)}{220} \;=\; \frac{6}{11} \tag{14}$$

We collect this information as in table 4.

TABLE 4. **Probabilities for Television Problem**

x	0	1	2
f(x)	6/11	9/22	1/22
F(x)	6/11	21/22	1

We compute the expected value as

$$\begin{aligned} E(X) \;&=\; \sum_{x\;\epsilon\;X} x\; p_X\,(x) \\ &= (0)\;\left(\frac{6}{11}\right) \;+\; (1)\;\left(\frac{9}{22}\right) \;+\; (2)\;\left(\frac{1}{22}\right) \;=\frac{11}{22} \;=\frac{1}{2} \end{aligned} \tag{15}$$

1.8. **Expected value of a function of a random variable.**

Theorem 1. *Let X be a discrete random variable with probability mass function p(x) and g(X) be a real-valued function of X. Then the expected value of g(X) is given by*

$$E[g(X)] \;=\; \sum_{x} \; g(x)\; p(x)\,. \tag{16}$$

Proof for case of finite values of X. Consider the case where the random variable X takes on a finite number of values $x_1, x_2, x_3, \cdots x_n$. The function g(x) may not be one-to-one (the different values of x_i may yield the same value of $g(x_i)$. Suppose that g(X) takes on m different values ($m \leq n$). It follows that g(X) is also a random variable with possible values $g_1, g_2, g_2, \ldots g_m$ and probability distribution

$$P[g(X) \;=\; g_i] \;=\; \sum_{\substack{\forall\, x_j such\; that \\ g(x_j)\,=\,g_i}} p(x_j) \;=\; p^*\,(g_i) \tag{17}$$

for all i = 1, 2, . . . m. Here $p*(g_i)$ is the probability that the experiment results in a value for the function f of the initial random variable of g_i. Using the definition of expected value in equation we obtain

$$E[g(X)] \;=\; \sum_{i=1}^{m} \; g_i \, p^*(g_i). \tag{18}$$

Now substitute in to obtain

$$\begin{aligned} E[g(X)] \;&=\; \sum_{i=1}^{m} \; g_i \, p^*(g_i)\,. \\ &= \sum_{i=1}^{m} \; g_i \left[\sum_{\substack{\forall\, x_j \,\ni \\ g\,(\,x_j\,) = g_i}} p\,(\,x_j\,) \right] \\ &= \sum_{i=1}^{m} \left[\sum_{\substack{\forall\, x_j \,\ni \\ g\,(\,x_j\,) = g_i}} g_i \; p\,(\,x_j\,) \right] \\ &= \sum_{j=1}^{n} \; g\,(x_j\,) \;\; p(\,x_j\,). \end{aligned} \tag{19}$$

□

1.9. **Properties of mathematical expectation.**

1.9.1. *Constants.*

Theorem 2. *Let X be a discrete random variable with probability function p(x) and c be a constant. Then E(c) = c.*

Proof. Consider the function g(X) = c. Then by theorem 1

$$E[c] \;\equiv\; \sum_{x} \; c\,p(x) \;=\; c \; \sum_{x} \; p(x) \tag{20}$$

But by property 1.4b, we have

$$\sum_{x} \; p(x) \;=\; 1$$

and hence

$$E\,(c) = c \cdot (1) = c. \tag{21}$$

□

1.9.2. *Constants multiplied by functions of random variables.*

Theorem 3. *Let X be a discrete random variable with probability function p(x), g(X) be a function of X, and let c be a constant. Then*

$$E\,[\,c\;g\,(\,X\,)\,]\;\equiv\;c\;E\;[\,(g\,(\,X\,)\,] \tag{22}$$

Proof. By theorem 1 we have

$$\begin{aligned} E[c\;g(X)]\;&\equiv\;\sum_{x}\;c\;g(x)\,p(x)\\ &=c\;\sum_{x}\;g(x)\;p(x)\\ &=c\;E[g(X)] \end{aligned} \tag{23}$$

□

1.9.3. *Sums of functions of random variables.*

Theorem 4. *Let X be a discrete random variable with probability function p(x),* $g_1(X), g_2(X), g_3(X), \cdots, g_k(X)$ *be k functions of X. Then*

$$E\,[g_1(X)+g_2(X)+g_3(X)+\cdots+g_k(X)]\;\equiv\;E[g_1(X)]+E[g_2(X)]+\cdots+E[g_k(X)] \tag{24}$$

Proof for the case of k = 2. By theorem 1 we have we have

$$\begin{aligned} E\,[g_1(X)\;+\;g_2\,(X)\,]\;&\equiv\;\sum_{x}\;[g_1\,(x)\;+\;g_2\,(x)\,]\;\,p(x)\\ &\equiv\;\sum_{x}\;g_1\,(x)\,p(x)\;+\;\sum_{x}\;g_2\,(x)\;\,p(x)\\ &=E\;[g_1\,(X)\,]\;+\;\;E\;[\,g_2\,(X)]\,, \end{aligned} \tag{25}$$

□

1.10. **Variance of a random variable.**

1.10.1. *Definition of variance.* The variance of a random variable X is defined to be the expected value of $(X-\mu)^2$. That is

$$V(X)\;=\;E\;\left[(\,X\;-\;\mu\,)^2\right] \tag{26}$$

The standard deviation of X is the positive square root of V(X).

1.10.2. *Example 1.* Consider a random variable with the following probability distribution.

TABLE 5. Probability Distribution for X

x	p(x)
0	1/8
1	1/4
2	3/8
3	1/4

We can compute the expected value as

$$\begin{aligned} \mu \;=\; E(X) \;&=\; \sum_{x=0}^{3} x\; p_X\left(x\right) \\ &= (0)\left(\frac{1}{8}\right) \;+\; (1)\left(\frac{1}{4}\right) \;+\; (2)\left(\frac{3}{8}\right) \;+\; (3)\left(\frac{1}{4}\right) \;=1\,\frac{3}{4} \end{aligned} \tag{27}$$

We compute the variance as

$$\begin{aligned} \sigma^2 \;&= E[X \;-\; \mu)^2] \;=\; \Sigma_{x=0}^{3}\,(x \;-\; \mu)^2\, p_X(x) \\ &= (0 \;-\; 1.75)^2\left(\frac{1}{8}\right) \;+\; (1 \;-\; 1.75)^2\left(\frac{1}{4}\right) \;+\; (2 \;-\; 1.75)^2\left(\frac{3}{8}\right) \;+\; (3 \;-\; 1.75)^2\left(\frac{1}{4}\right) \\ &= .9375 \end{aligned}$$

and the standard deviation as

$$\begin{aligned} \sigma^2 \;&= 0.9375 \\ \sigma \;&=\; +\,\sqrt{\sigma}^2 \;=\; \sqrt{.9375} \;=\; 0.97. \end{aligned}$$

1.10.3. *Alternative formula for the variance.*

Theorem 5. *Let X be a discrete random variable with probability function $p_X\left(x\right)$; then*

$$V\left(X\right) \;\equiv\; \sigma^2 \;=\; E\left[(X \;-\; \mu\,)^2\right] \;=\; E\left(X^2\right) \;-\; \mu^2 \tag{28}$$

Proof. First write out the first part of equation 28 as follows

$$\begin{aligned} V\left(X\right) \;\equiv\; \sigma^2 \;&=\; E\left[(X \;-\; \mu\,)^2\right] \;=\; E\left(X^2 \;-\; 2\,\mu\,X \;+\; \mu^2\right) \\ &= E\left(X^2\right) \;-\; E\left(2\,\mu\,X\right) \;+\; E\left(\mu^2\right) \end{aligned} \tag{29}$$

where the last step follows from theorem 4. Note that μ is a constant then apply theorems 3 and 2 to the second and third terms in equation 28 to obtain

$$V\left(X\right) \;\equiv\; \sigma^2 \;=\; E\left[(\,X \;-\; \mu\,)^2\,\right] \;=\; E\left(X^2\right) \;-\; 2\,\mu\,E\left(X\right) \;+\; \mu^2 \tag{30}$$

Then making the substitution that E(X) = μ, we obtain

$$V\left(X\right) \;\equiv\; \sigma^2 \;=\; E\left(X^2\right) \;-\; \mu^2 \tag{31}$$

□

1.10.4. *Example 2.* Die toss.

Toss a die that has six sides. Observe the number that comes up. The probability mass or frequency function is given by

$$p(x) \;=\; P(X \;=\; x) \;=\; \begin{cases} \frac{1}{6} & for\; x \;=\; 1,\, 2,\, 3,\, 4,\, 5,\, 6 \\ 0 & otherwise \end{cases}. \tag{32}$$

We compute the expected value as

$$\begin{aligned} E(X) &= \sum_{x \,\epsilon\, X} x\, p_X\,(x) \\ &= \sum_{i\,=\,1}^{6} i\ \left(\frac{1}{6}\right) \\ &= \frac{1+\ 2\ +\ 3\ +\ 4\ +\ 5\ +\ 6}{6} \\ &= \frac{21}{6}\ =\ 3\,\frac{1}{2} \end{aligned} \tag{33}$$

We compute the variance by then computing the $E(X^2)$ as follows

$$\begin{aligned} E(X^2) &= \sum_{x \,\epsilon\, X} x^2\ p_X\,(x) \\ &= \sum_{i\,=\,1}^{6} i^2\ \left(\frac{1}{6}\right) \\ &= \frac{1+\ 4\ +\ 9\ +\ 16\ +\ 2\ +\ 36}{6} \\ &= \frac{91}{6}\ =\ 15\,\frac{1}{6} \end{aligned} \tag{34}$$

We can then compute the variance using the formula Var(X) = E(X^2) - E^2(X) and the fact the E(X) = 21/6 from equation 33.

$$\begin{aligned} Var(X) &= E\,(X^2)\ -\ E^2(X) \\ &= \frac{91}{6}\ -\ \left(\frac{21}{6}\right)^2 \\ &= \frac{91}{6}\ -\ \left(\frac{441}{36}\right) \\ &= \frac{546}{36}\ -\frac{441}{36} \\ &= \frac{105}{36}\ =\ \frac{35}{12}\ =\ 2.916\overline{6} \end{aligned} \tag{35}$$

2. THE "DISTRIBUTION" OF RANDOM VARIABLES IN GENERAL

2.1. **Cumulative distribution function.** The cumulative distribution function (cdf) of a random variable X, denoted by $F_X(\cdot)$, is defined to be the function with domain the real line and range the interval [0,1], which satisfies $F_X(x) = P_X[X \leq x]$ = $P[\{\omega : X(\omega) \leq x\}]$ for every real number x. F has the following properties:

$$F_X(-\infty) = \lim_{x \to -\infty} F_X(x) = 0,\ F_X(+\infty) = \lim_{x \to +\infty} F_X(x) = 1, \tag{36a}$$

$$F_X(a) \leq F_X(b)\ for\ a < b,\ nondecreasing\ function\ of\ x, \tag{36b}$$

$$\lim_{0<h \to 0} F_X(x+h) = F_X(x),\ continuous\ from\ the\ right, \tag{36c}$$

2.2. **Example of a cumulative distribution function.** Consider the following function

$$F_X(x) = \frac{1}{1 + e^{-x}} \tag{37}$$

Check condition 36a as follows.

$$\begin{aligned} \lim_{x \to -\infty} F_X(x) &= \lim_{x \to -\infty} \frac{1}{1 + e^{-x}} = \lim_{x \to \infty} \frac{1}{1 + e^{x}} = 0 \\ \lim_{x \to \infty} F_X(x) &= \lim_{x \to \infty} \frac{1}{1 + e^{-x}} = 1 \end{aligned} \tag{38}$$

To check condition 36b differentiate the cdf as follows

$$\begin{aligned} \frac{d\,F_X(x)}{dx} &= \frac{d\left(\frac{1}{1+e^{-x}}\right)}{dx} \\ &= \frac{e^{-x}}{(1 + e^{-x})^2} > 0 \end{aligned} \tag{39}$$

Condition 36c is satisfied because $F_X(x)$ is a continuous function.

2.3. **Discrete and continuous random variables.**

2.3.1. *Discrete random variable.* A random variable X will be said to be discrete if the range of X is countable, that is if it can assume only a finite or countably infinite number of values. Alternatively, a random variable is discrete if $F_X(x)$ is a step function of x.

2.3.2. *Continuous random variable.* A random variable X is continuous if $F_X(x)$ is a continuous function of x.

2.4. **Frequency (probability mass) function of a discrete random variable.**

2.4.1. *Definition of a frequency (discrete density) function.* If X is a discrete random variable with the distinct values, $x_1, x_2, \cdots, x_n, \cdots$, then the function denoted by $p(\cdot)$ and defined by

$$p(x) = \begin{cases} P[X = x_j] & x = x_j,\ \ j = 1, 2, \ldots, n, \ldots \\ 0 & x \neq x_j \end{cases} \tag{40}$$

is defined to be the frequency, discrete density, or probability mass function of X. We will often write f(x) for p(x) to denote frequency as compared to probability.

A discrete probability distribution on R^k is a probability measure P such that

$$\sum_{i=1}^{\infty} P(\{x_i\}) = 1 \tag{41}$$

for some sequence of points in R^k, i.e. the sequence of points that occur as an outcome of the experiment. Given the definition of the frequency function in equation 40, we can also say that any non-negative function p on R^k that vanishes except on a sequence $x_1, x_2, \cdots, x_n, \cdots$ of vectors and that satisfies

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

defines a unique probability distribution by the relation

$$P(A) = \sum_{x_i \in A} p(x_i) \tag{42}$$

2.4.2. *Properties of discrete density functions.* As defined in section 1.4, a probability mass function must satisfy

$$p(x_j) > 0 \; for \; j = 1, 2, ... \tag{43a}$$

$$p(x) = 0 \; for x \neq x_j \; ; \; j = 1, 2, ..., \tag{43b}$$

$$\sum_j p(x)_j = 1 \tag{43c}$$

2.4.3. *Example 1 of a discrete density function.* Consider a probability model where there are two possible outcomes to a single action (say heads and tails) and consider repeating this action several times until one of the outcomes occurs. Let the random variable be the number of actions required to obtain a particular outcome (say heads). Let p be the probability that outcome is a head and (1-p) the probability of a tail. Then to obtain the first head on the xth toss, we need to have a tail on the previous x-1 tosses. So the probability of the first had occurring on the xth toss is given by

$$p(x) = P(X = x) = \begin{cases} (1-p)^{x-1} p & for \; x = 1, 2, ... \\ 0 & otherwise \end{cases} \tag{44}$$

For example the probability that it takes 4 tosses to get a head is 1/16 while the probability it takes 2 tosses is 1/4.

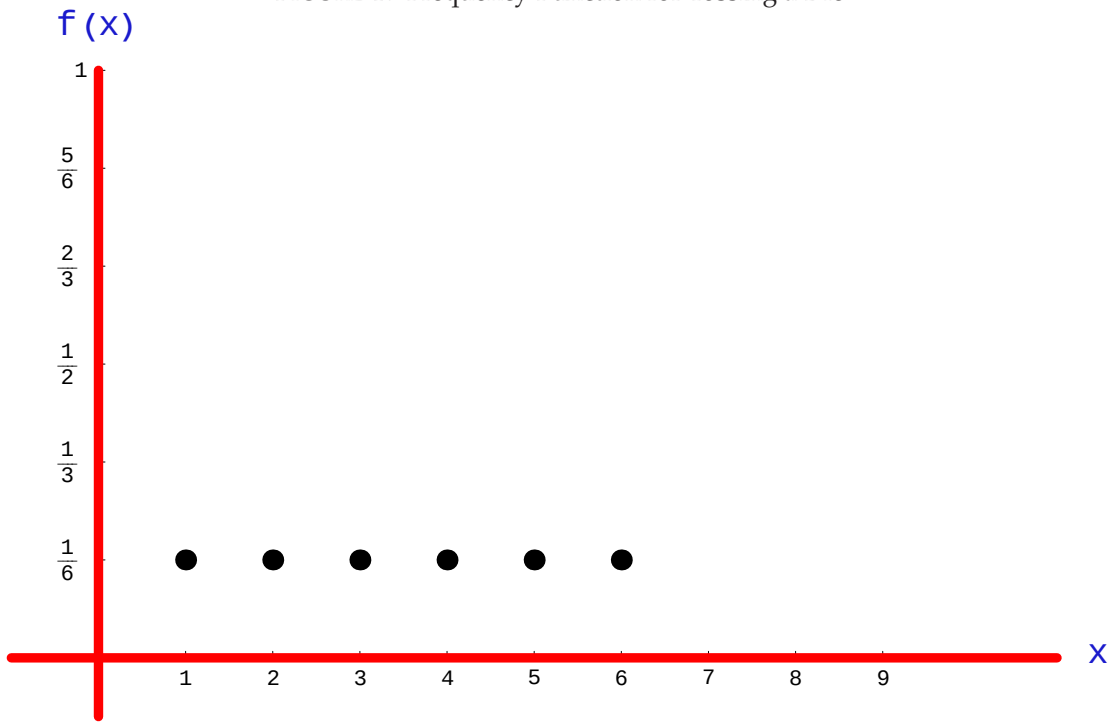
2.4.4. *Example 2 of a discrete density function.* Consider tossing a die. The sample space is {1, 2, 3, 4, 5, 6}. The elements are {1}, {2}, ... The frequency function is given by

$$p(x) = P(X = x) = \begin{cases} \frac{1}{6} & for \; x = 1, 2, 3, 4, 5, 6 \\ 0 & otherwise \end{cases} . \tag{45}$$

The density function is represented in figure 1.

2.5. **Probability density function of a continuous random variable.**

FIGURE 1. Frequency Function for Tossing a Die



2.5.1. *Alternative definition of continuous random variable.* In section 2.3.2, we defined a random variable to be continuous if $F_X(x)$ is a continuous function of x. We also say that a random variable X is continuous if there exists a function f(·) such that

$$F_X(x) = \int_{-\infty}^{x} f(u)\,du \tag{46}$$

for every real number x. The integral in equation 46 is a Riemann integral evaluated from -∞ to a real number x.

2.5.2. *Definition of a probability density frequency function (pdf).* The probability density function, $f_X(x)$, of a continuous random variable X is the function $f(\cdot)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du \tag{47}$$

2.5.3. *Properties of continuous density functions.*

$$f(x) \geq 0\ \forall x \tag{48a}$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1, \tag{48b}$$

Analogous to equation 42, we can write in the continuous case

$$P(X\ \epsilon\ A) = \int_{A} f_X(x)\ dx \tag{49}$$

where the integral is interpreted in the sense of Lebesgue.

Theorem 6. *For a density function f(x) defined over the set of all real numbers the following holds*

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx \tag{50}$$

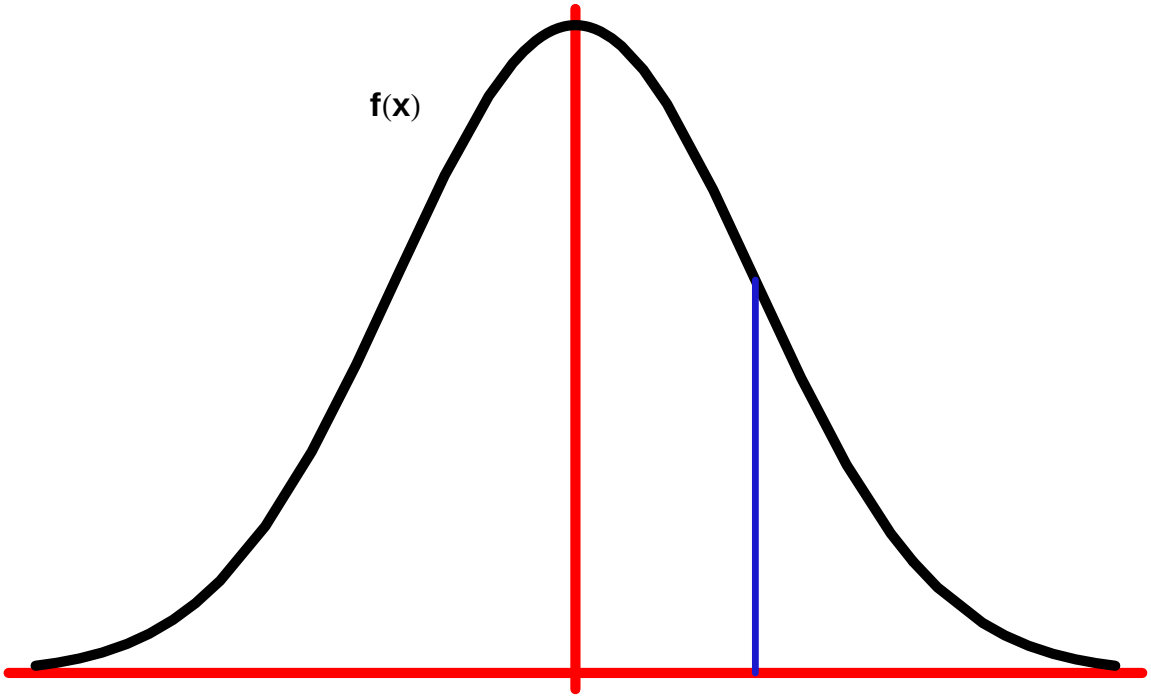
for any real constants a and b with $a \leq b$.

Also note that for a continuous random variable X the following are equivalent

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b) \tag{51}$$

Note that we can obtain the various probabilities by integrating the area under the density function as seen in figure 2.

FIGURE 2. Area under the Density Function as Probability



2.5.4. *Example 1 of a continuous density function.* Consider the following function

$$f(x) = \begin{cases} k \cdot e^{-3x} & for\ x > 0 \\ 0 & elsewhere \end{cases}. \tag{52}$$

First we must find the value of k that makes this a valid density function?
Given the condition in equation 48b we must have that

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_0^{\infty} k \cdot e^{-3x}\, dx = 1 \tag{53}$$

Integrate the second term to obtain

$$\int_0^\infty k \cdot e^{-3x}\, dx = k \cdot \lim_{t \to \infty} \frac{e^{-3x}}{-3} \Big|_0^t = \frac{k}{3} \tag{54}$$

Given that this must be equal to one we obtain

$$\begin{aligned} \frac{k}{3} &= 1 \\ \Rightarrow k &= 3 \end{aligned} \tag{55}$$

The density is then given by

$$f(x) = \begin{cases} 3 \cdot e^{-3x} & for\ x > 0 \\ 0 & elsewhere \end{cases}. \tag{56}$$

Now find the probability that (1 ≤ X ≤ 2).

$$\begin{aligned} P(1 \le X \le 2) &= \int_1^2 3 \cdot e^{-3x}\, dx \\ &= -e^{-3x}\Big|_1^2 \\ &= -e^{-6} + e^{-3} \\ &= -0.00247875 + 0.049787 \\ &= 0.047308 \end{aligned} \tag{57}$$

2.5.5. *Example 2 of a continuous density function.* Let X have p.d.f.

$$f(x) = \begin{cases} x \cdot e^{-x} & for\ x \le x \le \infty \\ 0 & elsewhere \end{cases}. \tag{58}$$

This density function is shown in figure 3.

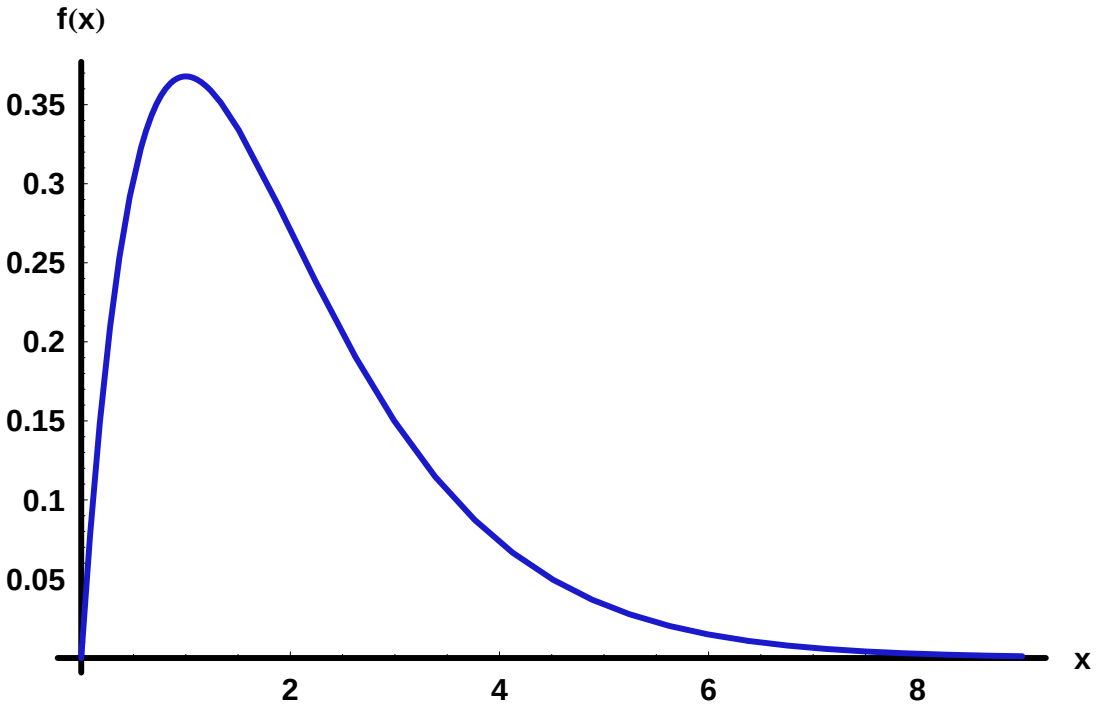
We can find the probability that (1 ≤ X ≤ 2) by integration

$$P(1 \le X \le 2) = \int_1^2 x \cdot e^{-x}\, dx \tag{59}$$

First integrate the expression on the right by parts letting u = x and dv = e^{-x} dx. Then du = dx and v = - e^{-x}dx. We then have

$$\begin{aligned} P(1 \le X \le 2) &= -x\,e^{-x}\Big|_1^2 - \int_1^2 -e^{-x}\, dx \\ &= -2\,e^{-2} + e^{-1} - \left[\, e^{-x}\,\Big|_1^2 \,\right] \\ &= -2\,e^{-2} + e^{-1} - e^{-2} + e^{-1} \\ &= -3\,e^{-2} + 2\,e^{-1} \\ &= \frac{-3}{e^2} + \frac{2}{e} \\ &= -0.406 + 0.73575 \\ &= 0.32975 \end{aligned} \tag{60}$$

FIGURE 3. Graph of Density Function $x\,e^{-x}$



This is represented by the area between the lines in figure 4.
We can also find the distribution function in this case.

$$F(x) = \int_0^x t \cdot e^{-t}\, dt \tag{61}$$

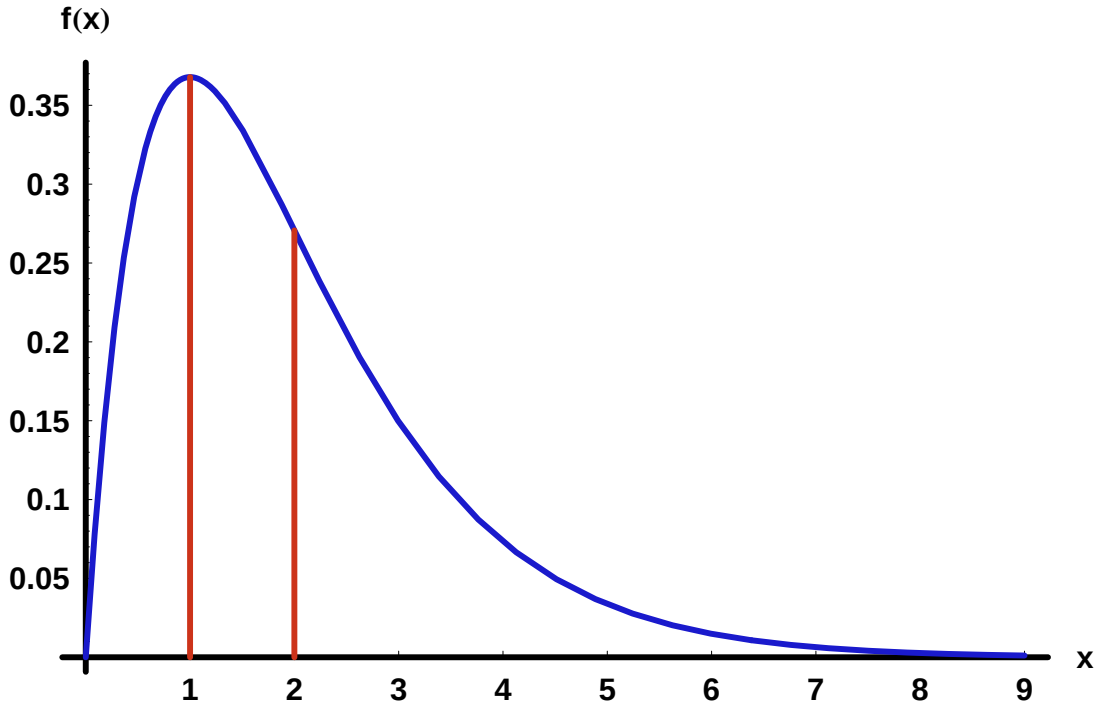
Make the u dv substitution as before to obtain

$$\begin{aligned} F(x) &= -t e^{-t}|_0^x - \int_0^x -e^{-t}\, dt \\ &= -t e^{-t}|_0^x - e^{-t}|_0^x \\ &= e^{-t}(-1 - t)|_0^x \\ &= e^{-x}(-1 - x) - e^{-0}(-1 - 0) \\ &= e^{-x}(-1 - x) + 1 \\ &= 1 - e^{-x}(1 + x) \end{aligned} \tag{62}$$

The distribution function is shown in figure 5.

Now consider the probability that (1 ≤ X ≤ 2)

FIGURE 4. $P(1 \leq X \leq 2)$



$$\begin{aligned} P(1 \leq X \leq 2) &= F(2) - F(1) \\ &= 1 - e^{-2}(1+2) - 1 + e^{-1}(1+1) \\ &= 2e^{-1} - 3e^{-2} \\ &= 0.73575 - 0.406 \\ &= 0.32975 \end{aligned} \tag{63}$$

We can see this as the difference in the values of F(x) at 1 and at 2 in figure 6

2.5.6. *Example 3 of a continuous density function.* Consider the normal density function given by

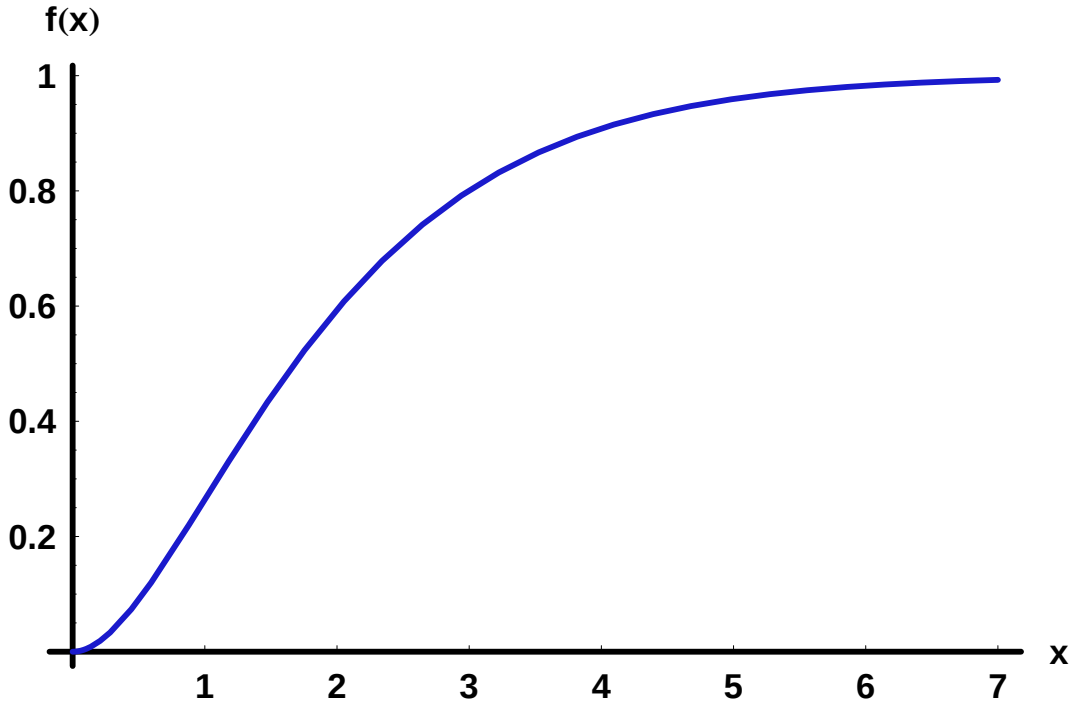
$$f(x : \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{64}$$

where μ and σ are parameters of the function. The shape and location of the density function depends on the parameters μ and σ. In figure 7 the diagram the density is drawn for $\mu = 0$, and $\sigma = 1$ and $\sigma = 2$.

2.5.7. *Example 4 of a continuous density function.* Consider a random variable with density function given by

$$f(x) = \begin{cases} (p+1)x^p & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases} \tag{65}$$

FIGURE 5. Graph of Distribution Function of Density Function $x\,e^{-x}$



where p is greater than -1. For example, if p = 0, then f(x) = 1, if p = 1, then f(x) = 2x and so on. The density function with p = 2 is shown in figure 8.

The distribution function with p = 2 is shown in figure 9.

2.6. **Expected value.**

2.6.1. *Expectation of a single random variable.* Let X be a random variable with density f(x). The expected value of the random variable, denoted E(X), is defined to be

$$E(X) = \begin{cases} \int_{-\infty}^{\infty} x\, f(x)\, dx & if\ X\ is\ continuous \\ \sum_{x\,\epsilon\,X} x\, p_X(x) & if\ X\ is\ discrete \end{cases}. \tag{66}$$

provided the sum or integral is defined. The expected value is kind of a weighted average. It is also sometimes referred to as the population mean of the random variable and denoted μ_X.

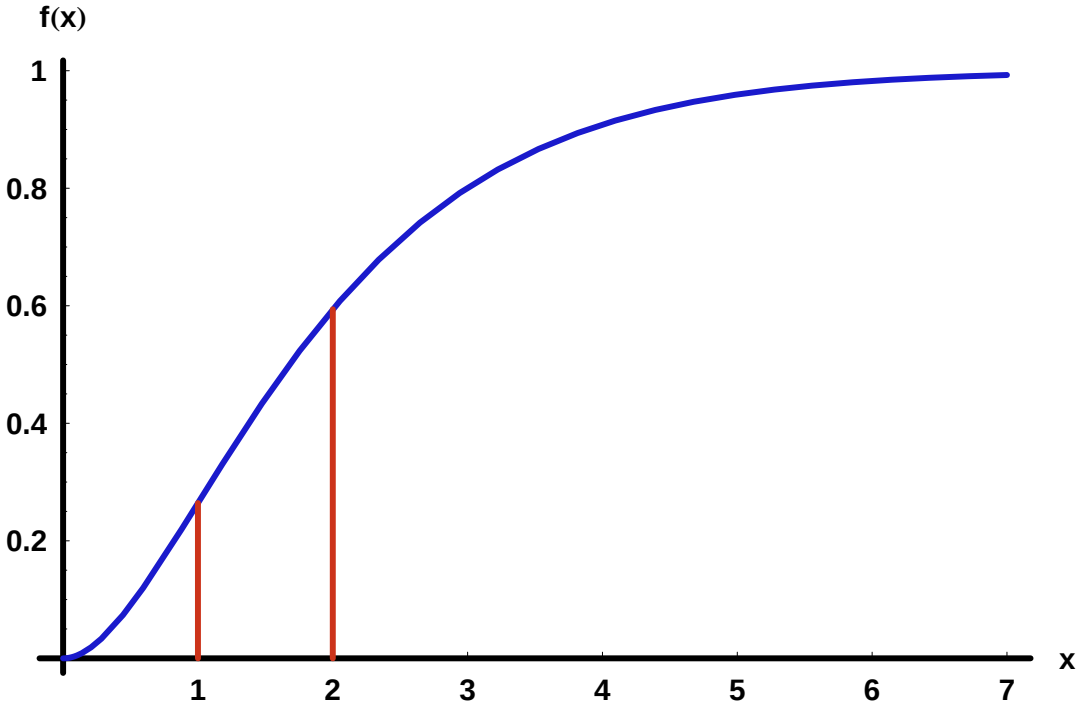
2.6.2. *Expectation of a function of a single random variable.* Let X be a random variable with density f(X). The expected value of a function g(·) of the random variable, denoted E(g(X)), is defined to be

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)\, f\,(x) dx \tag{67}$$

if the integral is defined.

The expectation of a random variable can also be defined using the Riemann-Stieltjes integral where F is a monotonically increasing function of X. Specifically

FIGURE 6. $P(1 \leq X \leq 2)$ using the Distribution Function



$$E(X) \;=\; \int_{-\infty}^{\infty} x\, dF(x) \;=\; \int_{-\infty}^{\infty} x\, dF \tag{68}$$

2.7. **Properties of expectation.**

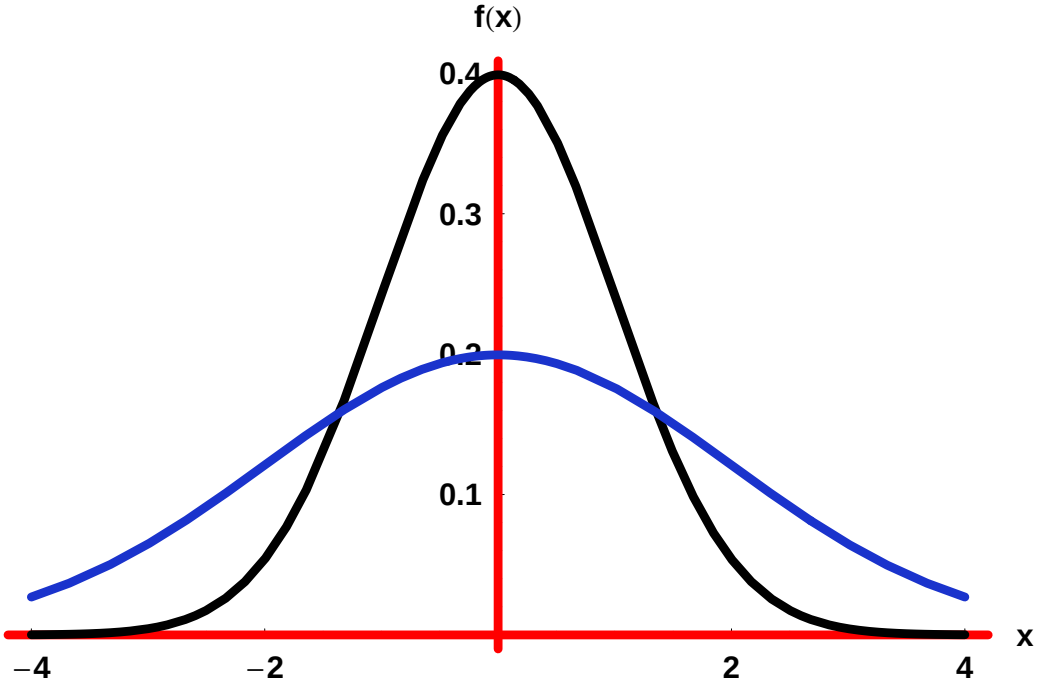
2.7.1. *Constants.*

$$\begin{aligned} E[a] \;&\equiv \int_{-\infty}^{\infty} a\, f(x) dx \\ &\equiv a \int_{-\infty}^{\infty} f(x) dx \\ &\equiv a \end{aligned} \tag{69}$$

2.7.2. *Constants multiplied by a random variable.*

$$\begin{aligned} E[a\, X] \;&\equiv \int_{-\infty}^{\infty} a\, x\, f(x) dx \\ &\equiv a \int_{-\infty}^{\infty} x\, f(x) dx \\ &\equiv a\, E[X] \end{aligned} \tag{70}$$

FIGURE 7. Normal Density Function



2.7.3. *Constants multiplied by a function of a random variable.*

$$\begin{aligned} E[a\,g(X)] &\equiv \int_{-\infty}^{\infty} a\ g(x)\ f(x)dx \\ &\equiv a\ \int_{-\infty}^{\infty} g(x)\,f(x)dx \\ &\equiv a\ E[g(X)] \end{aligned} \tag{71}$$

2.7.4. *Sums of expected values.* Let X be a continuous random variable with density function f(x) and let $g_1\,(X), g_2\,(X), g_3\,(X), \cdots, g_k\,(X)$ be k functions of X. Also let $c_1,\ c_2,\ c_3,\ \cdots c_k$ be k constants. Then

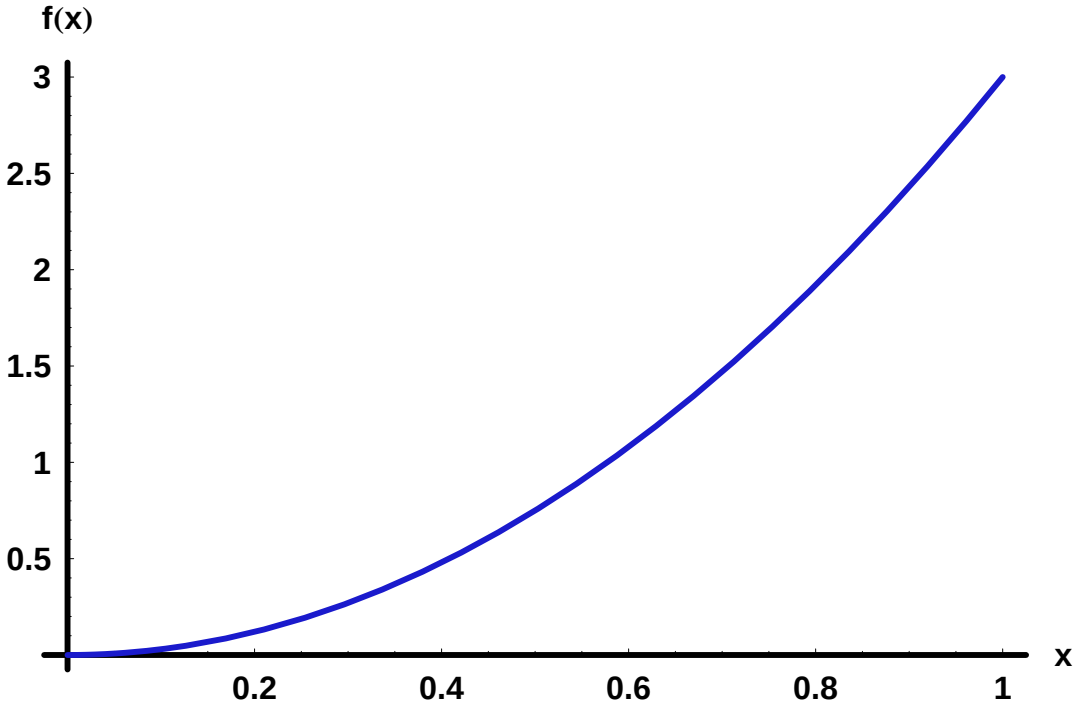
$$E\left[c_1\,g_1(X) + c_2\,g_2(X) + \cdots + c_k\,g_k(X)\ \right] \equiv\ E\left[c_1\,g_1(X)\right] + E\left[c_2\,g_2(X)\right]\ +\ \cdots\ +\ E\left[c_k\,g_k(X)\right] \tag{72}$$

2.8. **Example 1.** Consider the density function

$$f(x)\ = \begin{cases} (p+1)x^p & 0\ \leq\ x\ \leq\ 1 \\ 0 & otherwise \end{cases} \tag{73}$$

where p is greater than -1. We can compute the E(X) as follows.

FIGURE 8. Density Function $(p+1)\,x^p$



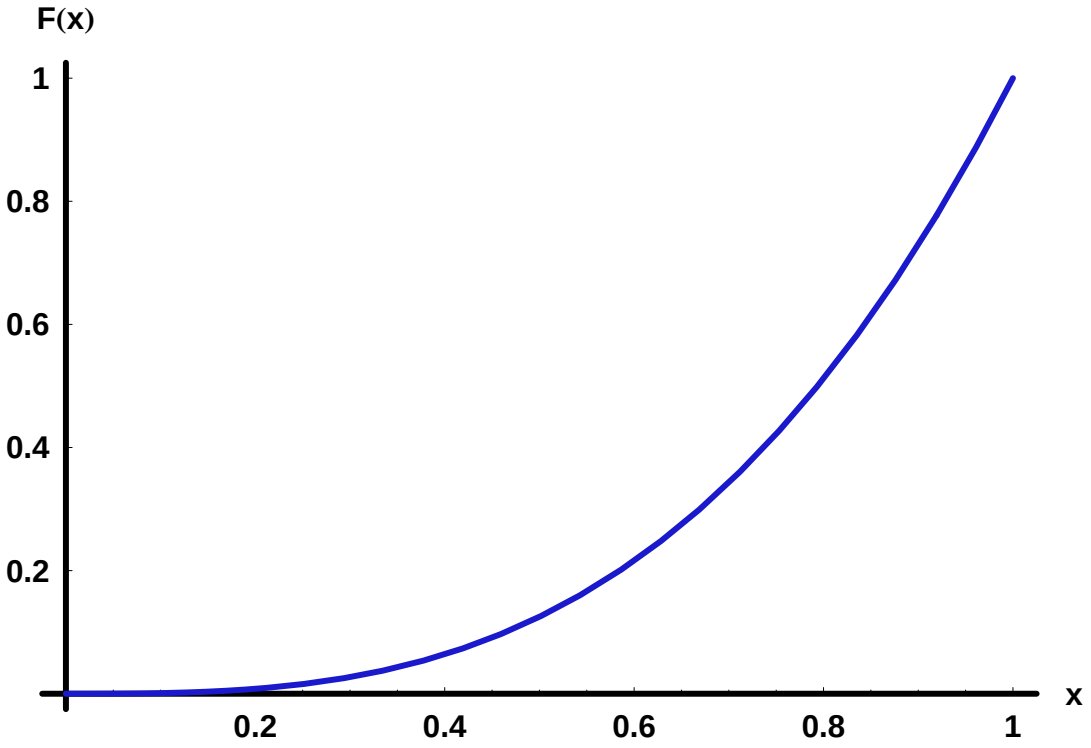
$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x)dx \\ &= \int_0^1 x(p+1)x^p dx \\ &= \int_0^1 x^{(p+1)}(p+1)dx \\ &= \frac{x^{(p+2)}(p+1)}{(p+2)} \Big|_0^1 \\ &= \frac{p+1}{p+2} \end{aligned} \tag{74}$$

2.9. **Example 2.** Consider the exponential distribution which has density function

$$f(x) = \frac{1}{\lambda}\, e^{\frac{-x}{\lambda}}\ 0 \leq x \leq \infty\,,\ \lambda > 0 \tag{75}$$

We can compute the E(X) as follows.

FIGURE 9. Density Function $(p = 1)\, x^p$



$$\begin{aligned} E(X) &= \int_0^\infty x \tfrac{1}{\lambda}\, e^{\frac{-x}{\lambda}}\, dx \\ &= -x\, e^{\frac{-x}{\lambda}}\, |_0^\infty + \int_0^\infty e^{\frac{-x}{\lambda}}\, dx \left(u = \tfrac{x}{\lambda},\, du = \tfrac{1}{\lambda}\, dx,\, v = -\lambda e^{\frac{-x}{\lambda}},\, dv = e^{\frac{-x}{\lambda}}\, dx \right) \\ &= 0 + \int_0^\infty e^{\frac{-x}{\lambda}} dx \\ &= -\lambda\, e^{\frac{-x}{\lambda}}\, |_0^\infty \\ &= \lambda \end{aligned} \tag{76}$$

2.10. **Variance.**

2.10.1. *Definition of variance.* The variance of a single random variable X with mean μ is given by

$$\begin{aligned} Var(X) \equiv \sigma^2 &\equiv E\left[(X - E(X))^2 \right] \\ &\equiv E\left[(X - \mu)^2 \right] \\ &\equiv \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \end{aligned} \tag{77}$$

We can write this in a different fashion by expanding the last term in equation 77.

$$
\begin{aligned}
Var(X) &\equiv \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \\
&\equiv \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)dx \\
&\equiv \int_{-\infty}^{\infty} x^2 f(x)\ dx - 2\mu \int_{-\infty}^{\infty} x\ f(x)\ dx + \mu^2 \int_{-\infty}^{\infty} f(x)\ dx \\
&= E\left[X^2\right] - 2\mu E\left[X\right] + \mu^2 \\
&= E\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= E\left[X^2\right] - \mu^2 \\
&\equiv \int_{-\infty}^{\infty} x^2 f(x)dx - \left[\int_{-\infty}^{\infty} x f(x)dx\right]^2
\end{aligned}
\tag{78}
$$

The variance is a measure of the dispersion of the random variable about the mean.

2.10.2. *Variance example 1.* Consider the density function

$$
f(x) = \begin{cases} (p+1)x^p & 0 \le x \le 1 \\ 0 & otherwise \end{cases}
\tag{79}
$$

where p is greater than -1. We can compute the Var(X) as follows.

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f(x)dx \\
&= \int_0^1 x(p+1)x^p dx \\
&= \frac{x^{(p+2)}(p+1)}{(p+2)}\Big|_0^1 \\
&= \frac{p+1}{p+2} \\
E(X^2) &= \int_0^1 x^2 (p+1)x^p\ dx \\
&= \frac{x^{(p+3)}(p+1)}{(p+3)}\Big|_0^1 \\
&= \frac{p+1}{p+3} \\
Var(X) &= E(X^2) - E^2(X) \\
&= \frac{p+1}{p+3} - \left(\frac{p+1}{p+2}\right)^2 \\
&= \frac{p+1}{(p+2)^2\ (p+3)}
\end{aligned}
\tag{80}
$$

The values of the mean and variances for various values of p are given in table 6.

TABLE 6. Mean and Variance for Distribution $f(x) = (p+1)\,x^p$ for alternative values of p

p	-.5	0	1	2	∞
E(x)	0.333	0.5	0.66667	0.75	1
Var(x)	0.08888	0.833333	0.277778	0.00047	0

2.10.3. *Variance example 2.* Consider the exponential distribution which has density function

$$f(x) = \frac{1}{\lambda}\, e^{\frac{-x}{\lambda}}\ 0 \le x \le \infty,\ \lambda > 0 \tag{81}$$

We can compute the $E(X^2)$ as follows

$$\begin{aligned}
E(X^2) &= \int_0^\infty x^2\, \frac{1}{\lambda}\, e^{\frac{-x}{\lambda}}\, dx \\
&= -x^2 e^{\frac{-x}{\lambda}}\,|_0^\infty + 2\int_0^\infty x\, e^{\frac{-x}{\lambda}}\, dx \ \left(u = \frac{x^2}{\lambda}, du = \frac{2\,x}{\lambda}\, dx, v = -\,\lambda\, e^{\frac{-x}{\lambda}}, dv = e^{\frac{-x}{\lambda}} dx\right) \\
&= 0 + 2\int_0^\infty x\, e^{\frac{-x}{\lambda}}\, dx \\
&= -\,2\,\lambda\, x\, e^{\frac{-x}{\lambda}}\,|_0^\infty + 2\int_0^\infty \lambda\, e^{\frac{-x}{\lambda}}\, dx \ \left(u = 2\,x,\ du = 2\,dx,\ v = -\,\lambda\, e^{\frac{-x}{\lambda}}, dv = e^{\frac{-x}{\lambda}} dx\right) \\
&= 0 + 2\,\lambda\int_0^\infty e^{\frac{-x}{\lambda}} dx \\
&= (2\,\lambda)\left(-\,\lambda\, e^{\frac{-x}{\lambda}}\,|_0^\infty\right) \\
&= (2\,\lambda)\,(\,\lambda\,) \\
&= 2\,\lambda^2
\end{aligned} \tag{82}$$

We can then compute the variance as

$$\begin{aligned}
Var(X) &= E\,(X^2) - E^2\,(X) \\
&= 2\,\lambda^2 - \lambda^2 \\
&= \lambda^2
\end{aligned} \tag{83}$$

3. MOMENTS AND MOMENT GENERATING FUNCTIONS

3.1. **Moments.**

3.1.1. *Moments about the origin (raw moments).* The rth moment about the origin of a random variable X, denoted by μ_r', is the expected value of X^r; symbolically,

$$\begin{aligned}
\mu_r' &= E(X^r) \\
&= \sum_x x^r f(x)
\end{aligned} \tag{84}$$

for r = 0, 1, 2, . . . when X is discrete and

$$\begin{aligned} \mu'_r &= E(\,X^r) \\ &= \int_{-\infty}^{\infty} x^r f(x)\,dx \end{aligned} \tag{85}$$

when X is continuous. The rth moment about the origin is only defined if $E[X^r]$ exists. A moment about the origin is sometimes called a raw moment. Note that $\mu'_1 = E(X) = \mu_X$, the mean of the distribution of X, or simply the mean of X. The rth moment is sometimes written as function of θ where θ is a vector of parameters that characterize the distribution of X.

3.1.2. *Central moments.* The rth moment about the mean of a random variable X, denoted by μ_r, is the expected value of $(X - \mu_X)^r$ symbolically,

$$\begin{aligned} \mu_r &= E[(X - \mu_X)^r] \\ &= \sum_x (x - \mu_X)^r f(x) \end{aligned} \tag{86}$$

for r = 0, 1, 2, . . . when X is discrete and

$$\begin{aligned} \mu_r &= E[(X - \mu_X)^r] \\ &= \int_{-\infty}^{\infty} (x - \mu_X)^r f(x)\,dx \end{aligned} \tag{87}$$

when X is continuous. The rth moment about the mean is only defined if $E[(X - \mu_X)^r]$ exists. The rth moment about the mean of a random variable X is sometimes called the rth central moment of X. The rth central moment of X about a is defined as $E[(X-a)^r]$. If a = μ_X, we have the rth central moment of X about μ_X. Note that $\mu_1 = E[(X - \mu_X)] = 0$ and $\mu_2 = E[(X - \mu_X)^2]$ = Var[X]. Also note that all odd moments of X around its mean are zero for symmetrical distributions, provided such moments exist.

3.1.3. *Alternative formula for the variance.*

Theorem 7.

$$\sigma_X^2 = \mu'_2 - \mu_X^2 \tag{88}$$

Proof.

$$\begin{aligned} Var(X) \equiv \sigma_X^2 &\equiv E\left[(X - E(X))^2\right] \\ &\equiv E\left[(X - \mu_X)^2\right] \\ &\equiv E\left[X^2 - 2\,\mu_X X + \mu_X^2\right] \\ &= E\left[X^2\right] - 2\,\mu_X E\,[X] + \mu^2 \\ &= E\left[X^2\right] - 2\,\mu_X^2 + \mu_X^2 \\ &= E\left[X^2\right] - \mu_X^2 \\ &= \mu'_2 - \mu_X^2 \end{aligned} \tag{89}$$

□

3.2. **Moment generating functions.**

3.2.1. *Definition of a moment generating function.* The moment generating function of a random variable X is given by

$$M_X(t) \;=\; E\, e^{t\,X} \tag{90}$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, $E\, e^{tX}$ exists. We can write $M_X(t)$ as

$$M_X(\,t\,) \;=\; \begin{cases} \int_{-\infty}^{\infty} e^{t\,x}\, f_X\,(x)\; dx & if\ X\ is\ continuous \\ \sum_x e^{t\,x}\, P\;(X \;=\; x) & if\ X\ is\ discrete \end{cases}. \tag{91}$$

To understand why we call this a moment generating function consider first the discrete case. We can write e^{tx} in an alternative way using a Maclaurin series expansion. The Maclaurin series of a function f(t) is given by

$$\begin{aligned} f(t) \;&=\; \sum_{n\,=\,0}^{\infty}\; \frac{f^{(n)}\,(0)}{n\,!}\; t^n \;=\; \sum_{n\,=\,0}^{\infty}\; f^{(n)}\,(0)\; \frac{t^n}{n\,!} \\ &= f\;(\,0\,) \;+\; \frac{f^{(1)}\;(0)}{1\,!}\; t \;+\; \frac{f\,^{(\,2\,)}\;(0)}{2\,!}\; t^2 \;+\; \frac{f^{(3)}\;(0)}{3\,!}\; t^3 \;+\; \cdots + \\ &= f\;(\,0\,) \;+\; f^{(1)}(0)\,\frac{t}{1\,!} \;+\; f\,^{(\,2\,)}(0)\,\frac{t^2}{2\,!} \;+\; f^{(3)}(0)\,\frac{t^3}{3\,!} \;+\; \cdots + \end{aligned} \tag{92}$$

where $f^{(n)}$ is the nth derivative of the function with respect to t and $f^{(n)}(0)$ is the nth derivative of f with respect to t evaluated at t = 0. For the function e^{tx}, the requisite derivatives are

$$\begin{aligned} \frac{d\,e^{t\,x}}{d\,t} \;&=\; x\,e^{t\,x}, \quad \left.\frac{d\,e^{t\,x}}{d\,t}\right]_{\,t\,=\,0} \;=\; x \\ \frac{d^2\,e^{t\,x}}{d\,t^2} \;&=\; x^2\,e^{t\,x}, \quad \left.\frac{d^2\,e^{t\,x}}{d\,t^2}\right]_{\,t\,=\,0} \;=\; x^2 \\ \frac{d^3\,e^{t\,x}}{d\,t^3} \;&=\; x^3\,e^{t\,x}, \quad \left.\frac{d^3\,e^{t\,x}}{d\,t^3}\right]_{\,t\,=\,0} \;=\; x^3 \\ &\qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots \\ \frac{d^j\,e^{t\,x}}{d\,t^j} \;&=\; x^j\,e^{t\,x}, \quad \left.\frac{d^j\;e^{t\,x}}{d\,t^j}\right]_{\,t\,=\,0} \;=\; x^j \end{aligned} \tag{93}$$

We can then write the Maclaurin series as

$$\begin{aligned} e^{t\,x} \;&=\; \sum_{n\,=\,0}^{\infty}\; \frac{d^n e^{t\,x}}{d\,t^n}\,(0)\; \frac{t^n}{n\,!} \\ &= \sum_{n\,=\,0}^{\infty}\; x^n\,\frac{t^n}{n\,!} \\ &= 1 \;+\; t\,x \;+\; \frac{t^2\,x^2}{2\,!} \;+\; \frac{t^3\,x^3}{3\,!} \;+\; \cdots + \;\frac{t^r x^r}{r\,!} \;+\; \cdots \end{aligned} \tag{94}$$

We can then compute $E(e^{tx}) = M_X\,(t)$ as

$$
\begin{aligned}
E\left[e^{t\,x}\right] &= M_X(t) \;=\; \sum_x e^{t\,x}\; f(x) \\
&= \sum_x \left[1 + t\,x + \;\frac{t^2\,x^2}{2\,!} + \frac{t^3\,x^3}{3!} \;+\; \cdots + \frac{t^{r\,x^r}}{r!} + \cdots\right]\; f(x) \\
&= \sum_x f(x) + t\sum_x x f(x) + \frac{t^2}{2!}\sum_x x^2 f(x) + \frac{t^3}{3!}\sum_x x^3 f(x) + \cdots + \frac{t^r}{r!}\sum_x x^r f(x) + \cdots \\
&= 1 + \mu t + \mu_2'\,\frac{t^2}{2!} + \mu_3'\,\frac{t^3}{3!} + \cdots + \mu_r'\frac{t^r}{r!} \;+\; \cdots
\end{aligned}
\tag{95}
$$

In the expansion, the coefficient of $\frac{t^r}{t!}$ is μ_r', the rth moment about the origin of the random variable X.

3.2.2. *Example derivation of a moment generating function.* Find the moment-generating function of the random variable whose probability density is given by

$$
f(x) \;= \begin{cases} e^{-x} & for\; x \;>\; 0 \\ 0 & elsewhere \end{cases}
\tag{96}
$$

and use it to find an expression for μ_r'. By definition

$$
\begin{aligned}
M_X(t) \;=\; E\left(e^{tX}\right) &= \int_{-\infty}^{\infty} e^{t\,x}\cdot e^{-x}\,dx \\
&= \int_o^{\infty} e^{-x\,(1\,-\,t)} dx \\
&= \frac{-1}{t \;-\; 1}\; e^{-x\,(1\,-\,t)}\;|_0^{\infty} \\
&= 0 \;-\; \left[\frac{-1}{1 \;-\; t}\right] \\
&= \frac{1}{1 \;-\; t}\; for\; t \;<\; 1
\end{aligned}
\tag{97}
$$

As is well known, when $|t| < 1$ the Maclaurin's series for $\frac{1}{1\,-\,t}$ is given by

$$
\begin{aligned}
M_x(t) \;= \frac{1}{1 \;-\; t} &= 1 \;+\; t \;+\; t^2 \;+\; t^3 \;+\; \cdots \;+\; t^r + \cdots \\
&= 1 \;+\; 1!\;\cdot\frac{t}{1!} \;+\; 2!\;\cdot\;\frac{t^2}{2!} \;+\; 3!\;\cdot\;\frac{t^3}{3}! \;+\; \cdots \;+\; r!\;\cdot\;\frac{t^r}{r!} \;+\; \cdots
\end{aligned}
\tag{98}
$$

or we can derive it directly using equation 92. To derive it directly utilizing the Maclaurin series we need the all derivatives of the function $\frac{1}{1\,-\,t}$ evaluated at 0. The derivatives are as follows

$$
\begin{aligned}
f(t) \;=\frac{1}{1\;-\;t} \;&=(1\;-\;t)^{-\,1}\\
f^{(1)} \;&=(1\;-\;t)^{-\,2}\\
f^{(2)} \;&=2\;(1\;-\;t)^{-3}\\
f^{(3)} \;&=6\;(1\;-\;t)^{-4}\\
f^{(4)} \;&=24\;(1\;-\;t)^{-5}\\
f^{(5)} \;&=120\;(1\;-\;t)^{-6}\\
&\vdots\\
f^{(n)} \;&=n\,!\;(1\;-\;t)^{(n\;+\;1)}\\
&\vdots
\end{aligned}
\tag{99}
$$

Evaluating them at zero gives

$$
\begin{aligned}
f(0) \;=\frac{1}{1\;-\;0} \;&=(1\;-\;0)^{-\,1}\;=\;1\\
f^{(1)} \;&=(1\;-\;0)^{-2}\;=\;1\;=\;1!\\
f^{(2)} \;&=2\;(1\;-\;0)^{-3}\;=\;2\;=\;2!\\
f^{(3)} \;&=6\;(1\;-\;0)^{-4}\;=\;6\;=\;3!\\
f^{(4)} \;&=24\;(1\;-\;0)^{-5}\;=\;24\;=\;4!\\
f^{(5)} \;&=120\;(1\;-\;0)^{-6}\;=\;120\;=\;5!\\
&\vdots\\
f^{(n)} \;&=n!\;(1\;-\;0)^{-\,(n\;+\;1)}\;=\;n!\\
&\vdots
\end{aligned}
\tag{100}
$$

Now substituting in appropriate values for the derivatives of the function $f(t)\;=\frac{1}{1\;-\;t}$ we obtain

$$
\begin{aligned}
f(t) \;&=\sum_{n\,=\,0}^{\infty}\;\frac{f^{(n)}\,(0)}{n!}\;t^{n}\\
&=f\,(0)\;+\;\frac{f^{(1)}\,(0)}{1!}\;t\;+\;\frac{f^{(2)}\,(0)}{2!}\;t^{2}\;+\;\frac{f^{(3)}\,(0)}{3!}\;t^{3}\;+\;\cdots\;+\\
&=1\;+\;\frac{1!}{1!}\;t\;+\;\frac{2!}{2!}\;t^{2}\;+\;\frac{3!}{3!}\;t^{3}\;+\;\cdots\;+\\
&=1\;+\;t\;+\;t^{2}\;+\;t^{3}\;+\;\cdots\;+
\end{aligned}
\tag{101}
$$

A further issue is to determine the radius of convergence for this particular function. Consider an arbitrary series where the nth term is denoted by a_n. The ratio test says that

$$If \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = L < 1\,,\ then\ the\ series\ is\ absolutely\ convergent \tag{102a}$$

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = L > 1\ or\ \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = \infty\,,\ then\ the\ series\ is\ divergent \tag{102b}$$

Now consider the nth term and the (n+1)th term of the Maclaurin series expansion of $\frac{1}{1-t}$.

$$\begin{aligned} a_n &= t^n \\ \lim_{n\to\infty}\left|\frac{t^{n+1}}{t^n}\right| &= \lim_{n\to\infty}|t| = L \end{aligned} \tag{103}$$

The only way for this to be less than one in absolute value is for the absolute value of t to be less than one, i.e., $|t| < 1$. Now writing out the Maclaurin series as in equation 98 and remembering that in the expansion, the coefficient of $\frac{t^r}{r!}$ is μ'_r, the rth moment about the origin of the random variable X

$$\begin{aligned} M_x(t) &= \frac{1}{1-t} = 1 + t + t^2 + t^3 + \cdots + t^r + \cdots \\ &= 1 + 1!\cdot\frac{t}{1!} + 2!\cdot\frac{t^2}{2!} + 3!\cdot\frac{t^3}{3}! + \cdots + r!\cdot\frac{t^r}{r!} + \cdots \end{aligned} \tag{104}$$

it is clear that μ'_r = r! for r = 0, 1, 2, ... For this density function E[X] = 1 because the coefficient of $\frac{t^1}{1!}$ is 1. We can verify this by finding E[X] directly by integrating.

$$E(X) = \int_0^\infty x\cdot e^{-x}\,dx \tag{105}$$

To do so we need to integrate by parts with u = x and dv = $e^{-x}dx$. Then du = dx and v = $-e^{-x}\,dx$. We then have

$$\begin{aligned} E(X) &= \int_0^\infty x\cdot e^{-x}\,dx,\ u = x,\ du = dx\,,\ v = -e^{-x}\,,\ dv = e^{-x}dx \\ &= -x\,e^{-x}\,|_0^\infty - \int_0^\infty -e^{-x}\,dx \\ &= [0 - 0] - \left[e^{-x}\,|_0^\infty\right] \\ &= 0 - [0 - 1] = 1 \end{aligned} \tag{106}$$

3.2.3. *Moment property of the moment generating functions for discrete random variables.*

Theorem 8. *If $M_X(t)$ exists, then for any positive integer k,*

$$\left.\frac{d^k M_X(t)}{dt^k}\right]_{t=0} = M_X^{(k)}(0) = \mu'_k\,. \tag{107}$$

In other words, if you find the kth derivative of M_X(t) with respect to t and then set t = 0, the result will be μ'_k .

Proof. $\frac{d^k M_X(t)}{dt^k}$, or $M_X^{(k)}(t)$, is the kth derivative of $M_X(t)$ with respect to t. From equation 95 we know that

$$M_X(t) \;=\; E\;\left(e^{t\,X}\right) \;=\; 1 \;+\; t\;\mu_1' \;+\; \frac{t^2}{2!}\;\mu_2' \;+\; \frac{t^3}{3!}\,\mu_3' \;+\; \cdots \tag{108}$$

It then follows that

$$M_X^{(1)}\,(t) \;=\mu_1' \;+\; \frac{2\,t}{2!}\;\mu_2' \;+\; \frac{3\,t^2}{3!}\;\mu_3' \;+\; \cdots \tag{109a}$$

$$M_X^{(2)}\,(t) \;=\mu_2' \;+\; \frac{2\,t}{2!}\;\mu_3' \;+\; \frac{3\,t^2}{3!}\;\mu_4' \;+\; \cdots \tag{109b}$$

where we note that $\frac{n}{n!} \;=\frac{1}{(n\;-\;1)\,!}$. In general we find that

$$M_X^{(\,k\,)}\,(\,t\,) \;=\; \mu_k' \;+\; \frac{2\,t}{2\,!}\;\mu_{k\,+\,1}' \;+\; \frac{3\,t^2}{3\,!}\;\mu_{k\,+\,2}' \;+\; \cdots\;. \tag{110}$$

Setting t = 0 in each of the above derivatives, we obtain

$$M_X^{(1)}\,(0) \;=\mu_1' \tag{111a}$$

$$M_X^{(2)}\,(0) \;=\mu_2' \tag{111b}$$

and, in general,

$$M_X^{(k)}\,(0) \;=\; \mu_k' \tag{112}$$

□

These operations involve interchanging derivatives and infinite sums, which can be justified if $M_X\,(t)$ exists.

3.2.4. *Moment property of the moment generating functions for continuous random variables.*

Theorem 9. *If X has mgf $M_X\,(t)$, then*

$$E\,X^n \;=\; M_X^{(n)}\;(0)\,, \tag{113}$$

where we define

$$M_X^{(n)}\,(0) \;=\frac{d^n}{dt^n}\;M_X\,(t)\,|_{\,t\,=\,0} \tag{114}$$

The nth moment of the distribution is equal to the nth derivative of $M_X\,(t)$ evaluated at t = 0.

Proof. We will assume that we can differentiate under the integral sign and differentiate equation 91.

$$\begin{aligned}
\frac{d}{dt}\;M_X\,(t) \;&=\frac{d}{dt}\;\int_{-\infty}^{\infty}\;e^{t\,x}\,f_X\,(\,x\,)\;dx \\
&=\int_{-\infty}^{\infty}\;\left(\frac{d}{dt}\;\;e^{t\,x}\right)\;f_X\,(\,x\,)\;dx \\
&=\int_{-\infty}^{\infty}\;\left(x\,e^{t\,x}\right)\;f_X\,(\,x\,)\;dx \\
&=E\;\left(X\,e^{\,t\,X}\right)
\end{aligned} \tag{115}$$

Now evaluate equation 115 at t = 0.

$$\frac{d}{dt}\, M_X\,(t)\,|_{t\,=\,0}\;=\;E\,\left(X\,e^{\,t\,X}\right)\,|t\;=\;0\;=\;E\;X \tag{116}$$

We can proceed in a similar fashion for other derivatives. We illustrate for n = 2.

$$\begin{aligned}\frac{d^2}{dt^2}\, M_X\,(t)\;&=\frac{d^2}{dt^2}\;\int_{-\infty}^{\infty}\,e^{t\,x}\,f_X\,(x\,)\;dx\\&=\int_{-\infty}^{\infty}\,\left(\frac{d^2}{dt^2}\;e^{t\,x}\right)\;f_X\,(x)\;dx\\&=\int_{-\infty}^{\infty}\,\left(\frac{d}{dt}\;x\,e^{t\,x}\right)\;f_X\,(x)\;dx\\&=\int_{-\infty}^{\infty}\,\left(x^2\,e^{t\,x}\right)\;f_X\,(x)\;dx\\&=E\,\left(\,X^2\,e^{\,t\,X}\,\right)\end{aligned} \tag{117}$$

Now evaluate equation 117 at t = 0.

$$\frac{d^2}{dt^2}\,M_X\,(t)\,|_{t\,=\,0}\;=\;E\left(X^2\,e^{\,t\,X}\right)\,|t\;=\;0\;=\;E\;X^2 \tag{118}$$

□

3.3. **Some properties of moment generating functions.** If a and b are constants, then

$$M_{X+a}\,(t)\;=E\,\left(e^{\,(X\,+\,a\,)\,t}\right)\;=\;e^{\,a\,t}\;\cdot\;M_X\;(t) \tag{119a}$$

$$M_{bX}\,(t)\;=E\,\left(e^{\,b\,X\,t}\right)\;=\;M_X(\,b\,t\,) \tag{119b}$$

$$M_{\frac{X\,+\,a}{b}}\;(t)\;=E\,\left(e^{\left(\frac{X+a}{b}\right)\,t}\right)\;=\;e^{\,\frac{a}{b}\,t}\;\cdot\;M_X\,\left(\frac{t}{b}\right) \tag{119c}$$

3.4. **Examples of moment generating functions.**

3.4.1. *Example 1.* Consider a random variable with two possible values, 0 and 1, and corresponding probabilities f(1) = p, f(0) = 1-p. For this distribution

$$\begin{aligned}M_X\,(t)\;&=E\,\left(e^{t\,X}\right)\\&=e^{\,t\,\cdot\,1}\,f\,(1)\;+\;e^{\,t\,\cdot\,0}\;f\,(0)\\&=e^{\,t}\,p\;+\;e^{\,0}\,(1\;-\;p)\\&=e^{\,0}\,(1\;-\;p)\;+\;e^{\,t}\,p\\&=1\;-\;p\;+\;e^{\,t}\,p\\&=1\;+\;p\,\left(e^{\,t}\;-\;1\right)\end{aligned} \tag{120}$$

The derivatives are

$$\begin{aligned} M_X^{(1)}(t) &= p\,e^t \\ M_X^{(2)}(t) &= p\,e^t \\ M_X^{(3)}(t) &= p\,e^t \\ &\vdots \\ M_X^{(k)}(t) &= p\,e^t \\ &\vdots \end{aligned} \tag{121}$$

Thus

$$E\left[X^k\right] \;=\; M_X^{(k)}(0) \;=\; p\,e^0 \;=\; p \tag{122}$$

We can also find this by expanding $M_X(t)$ using the Maclaurin series for the moment generating function for this problem

$$\begin{aligned} M_X(t) &= E\left(e^{t\,X}\right) \\ &= 1 \;+\; p\,\left(e^t \;-\; 1\right) \end{aligned} \tag{123}$$

To obtain this we first need the series expansion of e^t. All derivatives of e^t are equal to e^t. The expansion is then given by

$$\begin{aligned} e^t \;&=\; \sum_{n=0}^{\infty} \;\; \frac{d^n\,e^t}{d\,t^n}(0)\,\frac{t^n}{n!} \\ &= \sum_{n=0}^{\infty} \;\; \frac{t^n}{n!} \\ &= 1 \;+\; t \;+\; \frac{t^2}{2!} \;+\; \frac{t^3}{3!} \;+\; \cdots \;+\; \frac{t^r}{r!} \;+\; \cdots \end{aligned} \tag{124}$$

Substituting equation 124 into equation 123 we obtain

$$\begin{aligned} M_X(t) &= 1 \;+\; p\;e^t \;-\; p \\ &= 1 \;+\; p\;\left[1 \;+\; t \;+\; \frac{t^2}{2!} \;+\; \frac{t^3}{3!} \;+\; \cdots \;+\; \frac{t^r}{r\,!} \;+\; \cdots\right] \;-\; p \\ &= 1 \;+\; p \;+\; p\,t \;+\; p\,\frac{t^2}{2!} \;+\; p\,\frac{t^3}{3!} \;+\; \cdots \;+\; p\,\frac{t^r}{r!} \;+\; \cdots \;-\; p \\ &= 1 \;+\; p\,t \;+\; p\,\frac{t^2}{2!} \;+\; p\,\frac{t^3}{3!} \;+\; \cdots \;+\; p\,\frac{t^r}{r!} \;+\; \cdots \end{aligned} \tag{125}$$

We can then see that all moments are equal to p. This is also clear by direct computation

$$\begin{aligned} E(X) &= (1)\,p \;+\; (0)\,(1 \;-\; p) \;=\; p \\ E\left(X^2\right) &= (1^2)\,p \;+\; (0^2)\,(1 \;-\; p) \;=\; p \\ E\left(X^3\right) &= (1^3)\,p \;+\; (0^3)\,(1 \;-\; p) \;=\; p \\ &\vdots \\ E\left(X^k\right) &= (1^k)\,p \;+\; (0^k)\,(1 \;-\; p) \;=\; p \\ &\vdots \end{aligned} \tag{126}$$

3.4.2. *Example 2.* Consider the exponential distribution which has a density function given by

$$f(x) \;=\; \frac{1}{\lambda}\, e^{\frac{-x}{\lambda}}\; 0 \;\leq\; x \;\leq\; \infty\,,\; \lambda \;>\; 0 \tag{127}$$

For $\lambda\, t < 1$, we have

$$\begin{aligned} M_X(t) \;&=\; \int_0^\infty \; e^{t\,x}\; \frac{1}{\lambda}\; e^{\frac{-x}{\lambda}}\; dx \\ &=\frac{1}{\lambda}\; \int_0^\infty \; e^{-\left(\frac{1}{\lambda} \;-\; t\right)\,x}\; dx \\ &=\frac{1}{\lambda}\; \int_0^\infty \; e^{-\left(\frac{1 \;-\; \lambda\, t}{\lambda}\right)\,x}\; dx \\ &=\frac{1}{\lambda}\; \left[\frac{-\lambda}{1 \;-\; \lambda\, t}\right]\; e^{-\left(\frac{1 \;-\; \lambda\, t}{\lambda}\right)\,x}\,|_0^\infty \\ &= \left[\frac{-1}{1 \;-\; \lambda\, t}\right]\; e^{-\left(\frac{1 \;-\; \lambda\, t}{\lambda}\right)\,x}\,|_0^\infty \\ &= 0 \;-\; \left[\frac{-1}{1 \;-\; \lambda\, t}\right] e^0 \\ &=\frac{1}{1 \;-\; \lambda\, t} \end{aligned} \tag{128}$$

We can then find the moments by differentiation. The first moment is

$$\begin{aligned} E(X) \;&=\; \frac{d}{dt}\; (1 \;-\; \lambda\, t\,)^{\,-1}\;|_{t\,=\,0} \\ &= \lambda\,(1 \;-\; \lambda\, t)^{\,-2}\,|_{t\,=\,0} \\ &= \lambda \end{aligned} \tag{129}$$

The second moment is

$$\begin{aligned} E(X^2) &= \frac{d^2}{dt^2}\ (1\ -\ \lambda\, t)^{-1}\ |_{t=0} \\ &= \frac{d}{dt}\ \left(\lambda\ (1\ -\ \lambda\, t)^{-2}\right)\ |_{t=0} \\ &= 2\,\lambda^2\,(1\ -\ \lambda\, t)^{-3}\,|_{t=0} \\ &= 2\,\lambda^2 \end{aligned} \tag{130}$$

3.4.3. *Example 3.* Consider the normal distribution which has a density function given by

$$f(x\,;\ \mu,\,\sigma^2)\ =\ \frac{1}{\sqrt{2\pi\sigma^2}}\ \cdot\ e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{131}$$

Let g(x) = X - μ, where X is a normally distributed random variable with mean μ and variance σ^2. Find the moment-generating function for (X - μ). This is the moment generating function for central moments of the normal distribution.

$$M_X(t)\ =\ E[e^{t\,(X\ -\ \mu)}]\ =\ \frac{1}{\sqrt{2\pi\sigma^2}}\ \int_{-\infty}^{\infty}\ e^{t\ (\,x\,-\,\mu)}\ e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{132}$$

To integrate, let u = x - μ. Then du = dx and

$$\begin{aligned} M_X(t)\ =& \frac{1}{\sigma\sqrt{2\pi}}\ \int_{-\infty}^{\infty}\ e^{tu} e^{\frac{-u^2}{2\sigma^2}}\ du \\ =& \frac{1}{\sigma\sqrt{2\pi}}\ \int_{-\infty}^{\infty}\ e^{\left[t\,u\ -\ \frac{u^2}{2\,\sigma^2}\right]}\ du \\ =& \frac{1}{\sigma\sqrt{2\pi}}\ \int_{-\infty}^{\infty}\ e^{\left[\frac{1}{2\,\sigma^2}\ \left(2\,\sigma^2\,t\,u\ -\ u^2\right)\right]}\ du \\ =& \frac{1}{\sigma\sqrt{2\pi}}\ \int_{-\infty}^{\infty}\ \exp\ \left[\left(\frac{-1}{2\,\sigma^2}\right)\ (u^2\ -\ 2\sigma^2\,t\,u\,)\right]\ du \end{aligned} \tag{133}$$

To simplify the integral, complete the square in the exponent of e. That is, write the second term in brackets as

$$\left(u^2\ -\ 2\sigma^2\,t\,u\right)\ =\ \left(u^2\ -\ 2\sigma^2\,t\,u\ +\ \sigma^4\,t^2\ -\ \sigma^4\,t^2\right) \tag{134}$$

This then will give

$$\begin{aligned} \exp\left[\left(\frac{-1}{2\sigma^2}\right)(u^2-2\sigma^2tu)\right] =& \exp\left[\left(\frac{-1}{2\sigma^2}\right)(u^2-2\sigma^2tu+\sigma^4t^2-\sigma^4t^2)\right] \\ =& \exp\left[\left(\frac{-1}{2\,\sigma^2}\right)\ (u^2\ -\ 2\sigma^2\,t\,u+\sigma^4\,t^2)\right]\cdot\exp\ \left[\left(\frac{-1}{2\sigma^2}\right)(-\,\sigma^4t^2)\right] \\ =& \exp\ \left[\left(\frac{-1}{2\sigma^2}\right)(u^2-2\sigma^2tu+\sigma^4\,t^2)\right]\cdot\exp\left[\frac{\sigma^2t^2}{2}\right] \end{aligned} \tag{135}$$

Now substitute equation 135 into equation 133 and simplify. We begin by making the substitution and factoring out the term $\exp\left[\frac{\sigma^2t^2}{2}\right]$.

$$\begin{aligned} M_X(t) =& \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[\left(\frac{-1}{2\,\sigma^2}\right)\,(u^2 - 2\sigma^2\,t\,u)\right]\,du \\ =& \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[\left(\frac{-1}{2\,\sigma^2}\right)\,(u^2\,-\,2\sigma^2\,t\,u+\,\sigma^4\,t^2)\right]\cdot\exp\left[\frac{\sigma^2\,t^2}{2}\right]\,du \\ =& \exp\left[\frac{\sigma^2\,t^2}{2}\right]\,\left[\frac{1}{\sigma\sqrt{2\pi}}\right]\int_{-\infty}^{\infty}\exp\left[\left(\frac{-1}{2\,\sigma^2}\right)(u^2\,-\,2\sigma^2\,t\,u+\,\sigma^4\,t^2)\right]\,du \end{aligned} \tag{136}$$

Now move $\left[\frac{1}{\sigma\sqrt{2\pi}}\right]$ inside the integral sign, take the square root of $(u^2\,-\,2\sigma^2\,t\,u+\,\sigma^4\,t^2)$ and simplify

$$\begin{aligned} M_X(t) \;=\; & \exp\left[\frac{\sigma^2\,t^2}{2}\right]\,\int_{-\infty}^{\infty}\frac{\exp\,\left[\left(\frac{-1}{2\,\sigma^2}\right)\,(u^2-2\sigma^2\,t\,u+\sigma^4\,t^2)\right]}{\sigma\sqrt{2\pi}}\,du \\ =\; & \exp\left[\frac{\sigma^2\,t^2}{2}\right]\,\int_{-\infty}^{\infty}\frac{\exp\,\left[\left(\frac{-1}{2\,\sigma^2}\right)\,(u-\sigma^2\,t\,)^2\right]}{\sigma\sqrt{2\pi}}\,du \\ =\; & e^{\frac{t^2\,\sigma^2}{2}}\,\int_{-\infty}^{\infty}\frac{e^{\frac{-1}{2}\,\left[\frac{u-\sigma^2 t}{\sigma}\right]^2}}{\sigma\sqrt{2\pi}}\,du \end{aligned} \tag{137}$$

The function inside the integral is a normal density function with mean and variance equal to $\sigma^2 t$ and σ^2, respectively. Hence the integral is equal to 1. Then

$$M_X(t) \;=\; e^{\frac{t^2\,\sigma^2}{2}}\,. \tag{138}$$

The moments of u = x - μ can be obtained from M_X(t) by differentiating. For example the first central moment is

$$\begin{aligned} E(X\,-\,\mu\,) \;&=\frac{d}{dt}\,\left(e^{\frac{t^2\,\sigma^2}{2}}\right)\,|_{t=0} \\ &=t\;\sigma^2\;\left(e^{\frac{t^2\,\sigma^2}{2}}\right)\;|_{t=0} \\ &=0 \end{aligned} \tag{139}$$

The second central moment is

$$\begin{aligned} E(X\,-\,\mu\,)^2 \;&=\frac{d^2}{dt^2}\;\left(e^{\frac{t^2\,\sigma^2}{2}}\right)\;|_{t=0} \\ &=\frac{d}{dt}\;\left(t\;\sigma^2\;\left(e^{\frac{t^2\,\sigma^2}{2}}\right)\right)\;|_{t=0} \\ &=\left(t^2\,\sigma^4\;\left(e^{\frac{t^2\,\sigma^2}{2}}\right)\;+\;\sigma^2\;\left(e^{\frac{t^2\,\sigma^2}{2}}\right)\right)\;|_{t=0} \\ &=\sigma^2 \end{aligned} \tag{140}$$

The third central moment is

$$\begin{aligned} E(X - \mu)^3 &= \frac{d^3}{dt^3}\left(e^{\frac{t^2\sigma^2}{2}}\right)|_{t=0} \\ &= \frac{d}{dt}\left(t^2\sigma^4\left(e^{\frac{t^2\sigma^2}{2}}\right) + \sigma^2\left(e^{\frac{t^2\sigma^2}{2}}\right)\right)|_{t=0} \\ &= \left(t^3\sigma^6\left(e^{\frac{t^2\sigma^2}{2}}\right) + 2t\sigma^4\left(e^{\frac{t^2\sigma^2}{2}}\right) + t\sigma^4\left(e^{\frac{t^2\sigma^2}{2}}\right)\right)|_{t=0} \\ &= \left(t^3\sigma^6\left(e^{\frac{t^2\sigma^2}{2}}\right) + 3t\sigma^4\left(e^{\frac{t^2\sigma^2}{2}}\right)\right)|_{t=0} \\ &= 0 \end{aligned} \tag{141}$$

The fourth central moment is

$$\begin{aligned} E(X - \mu)^4 &= \frac{d^4}{dt^4}\left(e^{\frac{t^2\sigma^2}{2}}\right)|_{t=0} \\ &= \frac{d}{dt}\left(t^3\sigma^6\left(e^{\frac{t^2\sigma^2}{2}}\right) + 3t\sigma^4\left(e^{\frac{t^2\sigma^2}{2}}\right)\right)|_{t=0} \\ &= \left(t^4\sigma^8\left(e^{\frac{t^2\sigma^2}{2}}\right) + 3t^2\sigma^6\left(e^{\frac{t^2\sigma^2}{2}}\right) + 3t^2\sigma^6\left(e^{\frac{t^2\sigma^2}{2}}\right) + 3\sigma^4\left(e^{\frac{t^2\sigma^2}{2}}\right)\right)|_{t=0} \\ &= \left(t^4\sigma^8\left(e^{\frac{t^2\sigma^2}{2}}\right) + 6t^2\sigma^6\left(e^{\frac{t^2\sigma^2}{2}}\right) + 3\sigma^4\left(e^{\frac{t^2\sigma^2}{2}}\right)\right)|_{t=0} \\ &= 3\sigma^4 \end{aligned} \tag{142}$$

3.4.4. *Example 4.* Now consider the raw moments of the normal distribution. The density function is given by

$$f(x\,;\,\mu,\,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{143}$$

To find the moment-generating function for X we integrate the following function.

$$M_X(t) \;=\; E[e^{tX}] \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{tx}\, e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{144}$$

First rewrite the integral as follows by putting the exponents over a common denominator.

$$\begin{aligned} M_X(t) \;=\; E[e^{tX}] &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{tx}\, e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{\frac{-1}{2\sigma^2}(x-\mu)^2 + tx}\, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{\frac{-1}{2\sigma^2}(x-\mu)^2 + \frac{2\sigma^2 tx}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{\frac{-1}{2\sigma^2}\left[(x-\mu)^2 - 2\sigma^2 tx\right]} dx \end{aligned} \tag{145}$$

Now square the term in the exponent and simplify

$$\begin{aligned} M_X(t) \;=\; E[e^{tX}] \;&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \; e^{\frac{-1}{2\,\sigma^2}\left[x^2 \,-\, 2\,\mu\,x \,+\, \mu^2 \,-\, 2\,\sigma^2\,t\,x\right]} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \; e^{\frac{-1}{2\,\sigma^2}\left[\,x^2 \,-\, 2\,x\,\left(\,\mu \,+\, \sigma^2\,t\,\right) \,+\, \mu^2\,\right]} dx \end{aligned} \tag{146}$$

Now consider the exponent of e and complete the square for the portion in brackets as follows.

$$\begin{aligned} x^2 - 2x\,\left(\mu + \sigma^2\,t\right) + \mu^2 =& x^2 - 2\,x\,\left(\,\mu + \sigma^2\,t\right) + \mu^2 + \mathbf{2\mu\sigma^2 t} + \sigma^{\mathbf{4}}\mathbf{t^2} - 2\mu\sigma^2 t - \sigma^4 t^2 \\ &= \left(x^2 - (\mu + \sigma^2\,t)\right)^2 - 2\,\mu\,\sigma^2\,t - \sigma^4\,t^2 \end{aligned} \tag{147}$$

To simplify the integral, complete the square in the exponent of e by multiplying and dividing by

$$\left[e^{\frac{2\,\mu\,\sigma^2\,t + \sigma^4\,t^2}{2\,\sigma^2}}\right] \left[\,e^{\frac{-2\,\mu\,\sigma^2\,t \,-\, \sigma^4\,t^2}{2\,\sigma^2}}\right] \;=\; 1 \tag{148}$$

in the following manner

$$\begin{aligned} M_X\,(t) \;&=\; \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-1}{2\,\sigma^2}\left[x^2 - 2x\,\left(\mu + \sigma^2\,t\,\right) \,+\, \mu^2\right]} dx \\ &= \left[e^{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}}\right] \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-1}{2\sigma^2}\left[x^2 - 2x\left(\mu + \sigma^2 t\right) + \mu^2\right]} \left[e^{\frac{-\,2\mu\sigma^2 t - \sigma^4 t^2}{2\,\sigma^2}}\right] dx \\ &= \left[e^{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}}\right] \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-1}{2\sigma^2}\left[x^2 - 2x\left(\,\mu + \sigma^2 t\right) + \mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2\right]} dx \end{aligned} \tag{149}$$

Now find the square root of

$$x^2 - 2x\left(\mu + \sigma^2 t\right) + \mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 \tag{150}$$

Given we would like to have $(x - something)^2$, try squaring $x - (\mu + \sigma^2 t)$ as follows

$$\begin{aligned} \left[x - (\mu + \sigma^2 t)\right] \;&= x^2 \;-\; 2\left(x(\mu + \sigma^2 t)\right) \;+\; \left(\mu + \sigma^2 t\right)^2 \\ &= x^2 \;-\; 2x\left(\mu - \sigma^2 t\right) \;+\; \mu^2 \;+\; 2\mu\sigma^2 t \;+\; \sigma^4 t^2 \end{aligned} \tag{151}$$

So $\left[x - (\mu + \sigma^2 t)\right]$ is the square root of $x^2 \;-\; 2x\left(\mu - \sigma^2 t\right) \;+\; \mu^2 \;+\; 2\mu\sigma^2 t \;+\; \sigma^4 t^2$. Making the substitution in equation 149 we obtain

$$\begin{aligned} M_X\,(t) \;&=\; \left[e^{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}}\right] \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-1}{2\sigma^2}\left[x^2 - 2x\left(\mu + \sigma^2 t\right) + \mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2\right]} dx \\ &=\; \left[e^{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}}\right] \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-1}{2\sigma^2}\left(\left[x - (\mu + \sigma^2 t)\right]\right)} \end{aligned} \tag{152}$$

The expression to the right of $e^{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\,\sigma^2}}$ is a normal density function with mean and variance equal to $\mu + \sigma^2 t$ and σ^2, respectively. Hence the integral is equal to 1. Then

$$\begin{aligned} M_X(t) &= \left[e^{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}} \right] \\ &= e^{\mu t + \frac{t^2\sigma^2}{2}} \end{aligned} \tag{153}$$

The moments of X can be obtained from $M_X(t)$ by differentiating with respect to t. For example the first raw moment is

$$\begin{aligned} E(X) &= \frac{d}{dt}\left(e^{\mu\, t\, +\, \frac{t^2\sigma^2}{2}} \right) |_{t=0} \\ &= (\mu + t\,\sigma^2)\left(e^{\mu t + \frac{t^2\sigma^2}{2}} \right) |_{t=0} \\ &= \mu \end{aligned} \tag{154}$$

The second raw moment is

$$\begin{aligned} E(x^2) &= \frac{d^2}{dt^2}\left(e^{\mu\, t\, +\, \frac{t^2\sigma^2}{2}} \right) |_{t=0} \\ &= \frac{d}{dt}\left((\mu + t\,\sigma^2)\left(e^{\mu\, t + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &= \left(\left(\mu\ +\ t\ \sigma^2\right)^2 \left(e^{\mu\, t + \frac{t^2\sigma^2}{2}} \right) + \sigma^2 \left(e^{\mu t + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &= \mu^2\ +\ \sigma^2 \end{aligned} \tag{155}$$

The third raw moment is

$$\begin{aligned} E(X^3) &= \frac{d^3}{dt^3}\left(e^{\mu t + \frac{t^2\sigma^2}{2}} \right) |_{t=0} \\ &= \frac{d}{dt}\left(\left(\mu + t\sigma^2\right)^2 \left(e^{\mu t + \frac{t^2\sigma^2}{2}} \right) + \sigma^2 \left(e^{\mu t + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &= \left[\left(\mu + t\sigma^2\right)^3 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + 2\,\sigma^2 \left(\mu + t\sigma^2\right) \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + \sigma^2 \left(\mu + t\sigma^2\right) \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) \right] |_{t=0} \\ &= \left(\left(\mu + t\,\sigma^2\right)^3 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + 3\,\sigma^2 \left(\mu + t\sigma^2\right) \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &= \mu^3\ +\ 3\sigma^2\mu \end{aligned} \tag{156}$$

The fourth raw moment is

$$\begin{aligned} E(X^4) &= \frac{d^4}{dt^4}\left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) |_{t=0} \\ &= \frac{d}{dt}\left(\left(\mu + t\,\sigma^2\right)^3 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + 3\,\sigma^2 \left(\mu + t\,\sigma^2\right) \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &= \left(\left(\mu + t\,\sigma^2\right)^4 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + 3\sigma^2 \left(\mu + t\,\sigma^2\right)^2 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &+ \left(3\sigma^2 \left(\mu + t\,\sigma^2\right)^2 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + 3\,\sigma^4 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &= \left(\left(\mu + t\,\sigma^2\right)^4 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + 6\,\sigma^2 \left(\mu + t\,\sigma^2\right)^2 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) + 3\,\sigma^4 \left(e^{\mu + \frac{t^2\sigma^2}{2}} \right) \right) |_{t=0} \\ &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \end{aligned} \tag{157}$$

4. CHEBYSHEV'S INEQUALITY

Chebyshev's inequality applies equally well to discrete and continuous random variables. We state it here as a theorem.

4.1. **A Theorem of Chebyshev.**

Theorem 10. *Let X be a random variable with mean μ and finite variance σ^2. Then, for any constant $k > 0$,*

$$P(|X - \mu| < k\,\sigma) \geq 1 - \frac{1}{k^2} \qquad or \qquad P(|X - \mu| \geq k\,\sigma) \leq \frac{1}{k^2}\,. \tag{158}$$

The result applies for any probability distribution, whether the probability histogram is bell-shaped or not. The results of the theorem are very conservative in the sense that the actual probability that X is in the interval $\mu \pm k\sigma$ usually exceeds the lower bound for the probability, $1 - 1/k^2$, by a considerable amount.

Chebyshev's theorem enables us to find bounds for probabilities that ordinarily would have to be obtained by tedious mathematical manipulations (integration or summation). We often can obtain estimates of the means and variances of random variables without specifying the distribution of the variable. In situations like these, Chcbyshev's inequality provides meaningful bounds for probabilities of interest.

Proof. Let f(x) denote the density function of X. Then

$$\begin{aligned} V(X) = \sigma^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx \\ &= \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x)\,dx \\ &+ \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x)\,dx \\ &+ \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x)\,dx. \end{aligned} \tag{159}$$

The second integral is always greater than or equal to zero.

Now consider relationship between $(x - \mu)^2$ and $k\sigma^2$.

$$\begin{aligned} x &\leq \mu - k\sigma \\ \Rightarrow -x &\geq k\sigma - \mu \\ \Rightarrow \mu - x &\geq k\sigma \\ \Rightarrow (\mu - x)^2 &\geq k^2\sigma^2 \\ \Rightarrow (x - \mu)^2 &\geq k^2\sigma^2 \end{aligned} \tag{160}$$

And similarly,

$$\begin{aligned} x \;&\geq \mu \;+\; k\,\sigma \\ \Rightarrow\; x \;-\; \mu \;&\geq k\,\sigma \\ \Rightarrow\; (\,x \;-\; \mu\,)^{\,2} \;&\geq k^2\,\sigma^2 \end{aligned} \tag{161}$$

Now replace $(x - \mu)^2$ with $k\sigma^2$ in the first and third integrals of equation 159 to obtain the inequality

$$V(X) \;=\; \sigma^2 \;\geq\; \int_{-\infty}^{\mu\,-\,k\,\sigma} k^2\,\sigma^2\; f(x)\,dx \;+\; \int_{\mu\,+\,k\,\sigma}^{\infty} k^2\,\sigma^2\; f(x)\,dx\;. \tag{162}$$

Then

$$\sigma^2 \;\geq\; k^2\,\sigma^2\;\left[\int_{-\infty}^{\mu\,-\,k\,\sigma} f(x)\,dx \;+\; \int_{\mu\,+\,k\,\sigma}^{+\infty} f(x)\,dx\right] \tag{163}$$

W can write this in the following useful manner

$$\begin{aligned} \sigma^2 \;&\geq k^2\,\sigma^2\{P\,(\,X \;\leq\; \mu \;-\; k\,\sigma) \;+\; P\,(\,X+ \;\geq\; \mu \;+\; k\,\sigma)\} \\ &= k^2\,\sigma^2\,P\,(\,|\,X \;-\; \mu\,|\;\geq\; k\,\sigma). \end{aligned} \tag{164}$$

Dividing by $k^2\sigma^2$, we obtain

$$P\;(\,|\,X \;-\; \mu\,|\;\geq\; k\,\sigma\;) \leq \frac{1}{k^2}, \tag{165}$$

or, equivalently,

$$P\;(\,|\,X \;-\; \mu\,|\;<\; k\,\sigma\;) \;\geq\; 1 \;-\frac{1}{k^2}. \tag{166}$$

□

4.2. **Example.** The number of accidents that occur during a given month at a particular intersection, X, tabulated by a group of Boy Scouts over a long time period is found to have a mean of 12 and a standard deviation of 2. The underlying distribution is not known. What is the probability that, next month, X will be greater than eight but less than sixteen. We thus want $P[8 < X < 16]$. We can write equation 158 in the following useful manner.

$$P\;[(\mu \;-\; k\,\sigma\,) \;<\; X \;<\; (\,\mu \;+\; k\,\mu)] \;\geq\; 1 \;-\frac{1}{k^2} \tag{167}$$

For this problem μ = 12 and σ = 2 so $\mu - k\sigma$ = 12 - 2k. We can solve this equation for the k that gives us the desired bounds on the probability.

$$\begin{aligned} \mu \; - \; k\,\mu \; &= 12 - (k)\,(2) \; = \; 8 \\ \Rightarrow \; 2\,k \; &= 4 \\ \Rightarrow \; k \; &= 2 \\ & and \\ 12 \; + \; (\,k\,)\;(\,2\,) \; &= 16 \\ \Rightarrow \; 2\,k - 4 \\ \Rightarrow \; k \; &= 2 \end{aligned} \tag{168}$$

We then obtain

$$P\;[(8) \; < \; X \; < \; (16)\,] \; \geq \; 1 \; - \frac{1}{2^2} \; = \; 1 \; - \; \frac{1}{4} \; = \frac{3}{4} \tag{169}$$

Therefore the probability that X is between 8 and 16 is at least $3/4$.

4.3. **Alternative statement of Chebyshev's inequality.**

Theorem 11. *Let X be a random variable and let g(x) be a non-negative function. Then for* $r > 0$,

$$P\,[g(X) \; \geq \; r] \; \leq \; \frac{E\,g\,(X)}{r} \tag{170}$$

Proof.

$$\begin{aligned} E\,g\,(X) \; &= \int_{-\,\infty}^{\infty} \; g\,(x) \; f_X\,(x) \; dx \\ &\geq \int_{[x\,:\;g\,(x) \; \geq \; r]} \; g\;(x) \; f_X\,(x) \; dx \qquad\qquad (g \; is \; nonnegative) \\ &\geq r \int_{[\,x\,:\;g(x) \; \geq \; r\,]} \; f_X\,(x) \; dx \qquad\qquad (g\,(x) \; \geq \; r) \\ &= r\,P\;[\,g\,(X) \; \geq \; r\;] \\ \Rightarrow \; P\;[g\,(X) \; \geq \; r\,] \; &\leq \; \frac{E\,g\,(X)}{r} \end{aligned} \tag{171}$$

□

4.4. **Another version of Chebyshev's inequality as special case of general version.**

Corollary 1. Let X be a random variable with mean μ and variance σ^2. Then for any $k > 0$ or any $\varepsilon > 0$

$$P\;[\,|X \; - \; \mu| \; \geq \; k\,\sigma\,] \; \leq \; \frac{1}{k^2} \tag{172a}$$

$$P\;[\,|X \; - \; \mu| \; \geq \; \varepsilon\,] \; \leq \; \frac{\sigma^2}{\varepsilon^2} \tag{172b}$$

Proof. Let g(x) = $\frac{(x-\mu)^2}{\sigma^2}$, where μ = E(X) and σ^2 = Var (X). Then let r = k^2. Then

$$\begin{aligned} P\left[\frac{(X-\mu)^2}{\sigma^2} \geq k^2\right] &\leq \frac{1}{k^2} E\left(\frac{(X-\mu)^2}{\sigma^2}\right) \\ &= \frac{1}{k^2}\frac{E(X-\mu)^2}{\sigma^2} = \frac{1}{k^2} \end{aligned} \tag{173}$$

because $E(X-\mu)^2 = \sigma^2$. We can then rewrite equation 173 as follows

$$\begin{aligned} P\left[\frac{(X-\mu)^2}{\sigma^2} \geq k^2\right] &\leq \frac{1}{k^2} \\ \Rightarrow P\left[(X-\mu)^2 \geq k^2\sigma^2\right] &\leq \frac{1}{k^2} \\ \Rightarrow P\left[|X-\mu| \geq k\sigma\right] &\leq \frac{1}{k^2} \end{aligned} \tag{174}$$

□

REFERENCES

[1] Amemiya, T. *Advanced Econometrics.* Cambridge: Harvard University Press, 1985.
[2] Bickel P.J., and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol 1).* 2^{nd} Edition. Upper Saddle River, NJ: Prentice Hall, 2001.
[3] Billingsley, P. *Probability and Measure.* 3rd edition. New York: Wiley, 1995.
[4] Casella, G. And R.L. Berger. *Statistical Inference*. Pacific Grove, CA: Duxbury, 2002.
[5] Cramer, H. *Mathematical Methods of Statistics.* Princeton: Princeton University Press, 1946.
[6] Goldberger, A.S. *Econometric Theory.* New York: Wiley, 1964.
[7] Lindgren, B.W. *Statistical Theory* 4^{th} edition. Boca Raton, FL: Chapman & Hall/CRC, 1993.
[8] Rao, C.R. *Linear Statistical Inference and its Applications.* 2nd edition. New York: Wiley, 1973.
[9] Theil, H. *Principles of Econometrics*. New York: Wiley, 1971.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – II –TESTING OF HYPOTHESES AND DESIGN OF EXPERIMENTS – SMTA5205

Sampling theory is a study of relationships existing between a population and samples drawn from the population. Sampling theory is applicable only to random samples. For this purpose the population or a universe may be defined as an aggregate of items possessing a common trait or traits. In other words, a universe is the complete group of items about which knowledge is sought. The universe may be finite or infinite. Finite universe is one which has a definite and certain number of items, but when the number of items is uncertain and infinite, the universe is said to be an infinite universe. Similarly, the universe may be hypothetical or existent. In the former case the universe in fact does not exist and we can only imagine the items constituting it. Tossing of a coin or throwing a dice are examples of hypothetical universe. Existent universe is a universe of concrete objects i.e., the universe where the items constituting it really exist. On the other hand, the term sample refers to that part of the universe which is selected for the purpose of investigation. The theory of sampling studies the relationships that exist between the universe and the sample or samples drawn from it. The main problem of sampling theory is the problem of relationship between a parameter and a statistic. The theory of sampling is concerned with estimating the properties of the population from those of the sample and also with gauging the precision of the estimate. This sort of movement from particular (sample) towards general (universe) is what is known as statistical induction or statistical inference. In more clear terms "from the sample we attempt to draw inference concerning the universe. In order to be able to follow this inductive method, we first follow a deductive argument which is that we imagine a population or universe (finite or infinite) and investigate the behavior of the samples drawn from this universe applying the laws of probability.The methodology dealing with all this is known as sampling theory.

(*i*) *Statistical estimation:* Sampling theory helps in estimating unknown population parameters from a knowledge of statistical measures based on sample studies. In other words, to obtain an estimate of parameter from statistic is the main objective of the sampling theory. The estimate can either be a point estimate or it may be an interval estimate. Point estimate is a single estimate expressed in the form of a single figure, but interval estimate has two limits viz., the upper limit and the lower limit within which the parameter value may lie. Interval estimates are often used in statistical induction.

(*ii*) *Testing of hypotheses:* The second objective of sampling theory is to enable us to decide whether to accept or reject hypothesis; the sampling theory helps in determining whether observed differences are actually due to chance or whether they are really significant.

(*iii*) *Statistical inference:* Sampling theory helps in making generalisation about the population/ universe from the studies based on samples drawn from it. It also helps in determining the accuracy of such generalisations.

WHAT IS A HYPOTHESIS?

Ordinarily, when one talks about hypothesis, one simply means a mere assumption or some supposition to be proved or disproved. But for a researcher hypothesis is a formal question that he intends to resolve. Thus a hypothesis may be defined as a proposition or a set of

proposition set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional njecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable. For example, consider statements like the following ones: "Students who receive counselling will show a greater increase in creativity than students not receiving counselling" Or "the automobile A is performing as well as automobile B." These are hypotheses capable of being objectively verified and tested. Thus, we may conclude that a hypothesis states what we are looking for and it is a proposition which can be put to a test to determine its validity.

Characteristics of hypothesis: Hypothesis must possess the following characteristics:
(i) Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.
(ii) Hypothesis should be capable of being tested. In a swamp of untestable hypotheses, many a time the research programmes have bogged down. Some prior study may be done by researcher in order to make hypothesis a testable one. A hypothesis "is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation."
(iii) Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.
(iv) Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.
(v) Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.
(vi) Hypothesis should be consistent with most known facts i.e., it must be consistent with a substantial body of established facts. In other words, it should be one which judges accept as being the most likely.
(vii) Hypothesis should be amenable to testing within a reasonable time. One should not use even an excellent hypothesis, if the same cannot be tested in reasonable time for one cannot spend a life-time collecting data to test it.
(viii) Hypothesis must explain the facts that gave rise to the need for explanation. This means that by using the hypothesis plus other known and accepted generalizations, one should be able to deduce the original problem condition. Thus hypothesis must actually explain what it claims to explain; it should have empirical reference.

Basic concepts in the context of testing of hypotheses need to be explained.

(a) *Null hypothesis and alternative hypothesis:* In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. If we are to compare method *A* with method *B* about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis. As against this, we may think that the method *A* is superior or the method *B* is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H_0 and the alternative hypothesis as H_a. Suppose we want to test the hypothesis that the population mean (μ) is equal to the hypothesised mean $(\mu_{H_0}) = 100$.

Then we would say that the null hypothesis is that the population mean is equal to the hypothesised mean 100 and symbolically we can express as:

$$H_0 : \mu = \mu_{H_0} = 100$$

If our sample results do not support this null hypothesis, we should conclude that something else is true. What we conclude rejecting the null hypothesis is known as alternative hypothesis. In other words, the set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If we accept H_0, then we are rejecting H_a and if we reject H_0, then we are accepting H_a. For $H_0 : \mu = \mu_{H_0} = 100$, we may consider three possible alternative hypotheses as follows*:

Alternative hypothesis	*To be read as follows*
H_a : $_{H_0}$	(The alternative hypothesis is that the population mean is not equal to 100 i.e., it may be more or less than 100)
H_a : $_{H_0}$	(The alternative hypothesis is that the population mean is greater than 100)
H_a : $_{H_0}$	(The alternative hypothesis is that the population mean is less than 100)

The null hypothesis and the alternative hypothesis are chosen before the sample is drawn (the researcher must avoid the error of deriving hypotheses from the data that he collects and then testing the hypotheses from the same data). In the choice of null hypothesis, the following considerations are usually kept in view:

(a) Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject, and alternative hypothesis represents all other possibilities.

(b) If the rejection of a certain hypothesis when it is actually true involves great risk, it is taken as null hypothesis because then the probability of rejecting it when it is true is α (the level of significance) which is chosen very small.

(c) Null hypothesis should always be specific hypothesis i.e., it should not state about or approximately a certain value.

Generally, in hypothesis testing we proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Why so? The answer is that on the assumption that null hypothesis is true, one can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the alternative hypothesis. Hence the use of null hypothesis (at times also known as statistical hypothesis) is quite frequent.

(b) *The level of significance:* This is a very important concept in the context of hypothesis testing. It is always some percentage (usually 5%) which should be chosen with great care, thought and reason. In case we take the significance level at 5 per cent, then this implies that H_0 will be rejected

*If a hypothesis is of the type $\mu = \mu_{H_0}$, then we call such a hypothesis as simple (or specific) hypothesis but if it is of the type $\mu \neq \mu_{H_0}$ or $\mu > \mu_{H_0}$ or $\mu < \mu_{H_0}$, then we call it a composite (or nonspecific) hypothesis.

when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if H_0 is true. In other words, the 5 per cent level of significance means that researcher is willing to take as much as a 5 per cent risk of rejecting the null hypothesis when it (H_0) happens to be true. Thus the significance level is the maximum value of the probability of rejecting H_0 when it is true and is usually determined in advance before testing the hypothesis.

(c) *Decision rule or test of hypothesis:* Given a hypothesis H_0 and an alternative hypothesis H_a,

we make a rule which is known as decision rule according to which we accept H_0 (i.e., reject H_a) or reject H_0 (i.e., accept H_a). For instance, if (H_0 is that a certain lot is good (there are very few defective items in it) against H_a) that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis. We might test 10 items in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept H_0 otherwise we will reject H_0 (or accept H_a). This sort of basis is known as decision rule.

(d) *Type I and Type II errors:* In the context of testing of hypotheses, there are basically two types of errors we can make. We may reject H_0 when H_0 is true and we may accept H_0 when in fact H_0 is not true. The former is known as Type I error and the latter as Type II error. In other words, Type I error means rejection of hypothesis which should have been accepted and Type II error means accepting the hypothesis which should have been rejected. Type I error is denoted by α (alpha) known as α error, also called the level of significance of test; and Type II error is denoted by β (beta) known as β error. In a tabular form the said two errors can be presented as follows:

	Decision	
	Accept *H*	Reject *H*
H_0 (true)	Correct decision	Type I error (error)
H_0 (false)	Type II error (error)	Correct decision

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true. We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01.

But with a fixed sample size, *n*, when we try to reduce Type I error, the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously. There is a trade-off between two types of errors which means that the probability of making one type of error can only be reduced if we are willing to increase the probability of making the other type of error. To deal with this trade-off in business situations, decision-makers decide the appropriate level of Type I error by examining the costs or penalties attached to both types of errors. If Type I error involves the time and trouble of reworking a batch of chemicals that should have been accepted, whereas Type II error means taking a chance that an entire group of users of this chemical compound will be poisoned, then in such a situation one should prefer a Type I error to a Type II error. As a result one must set very high level for Type I error in one's testing technique of a given hypothesis.[2] Hence, in the testing of hypothesis, one must make all possible effort to strike an adequate balance between Type I and Type II errors.

(e) *Two-tailed and One-tailed tests:* In the context of hypothesis testing, these two terms are quite important and must be clearly understood. A two-tailed test rejects the null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesised value of the mean of the population. Such a test is appropriate when the null hypothesis is some specified value and the alternative hypothesis is a value not equal to the specified value of the null hypothesis.

TESTS OF HYPOTHESES

As has been stated above that hypothesis testing determines the validity of the assumption (technically described as null hypothesis) with a view to choose between two conflicting hypotheses about the value of a population parameter. Hypothesis testing helps to decide on the basis of a sample data, whether a hypothesis about the population is likely to be true or false. Statisticians have developed several tests of hypotheses (also known as the tests of significance) for the purpose of testing of hypotheses which can be classified as: (a) Parametric tests or standard tests of hypotheses; and

(b) Non-parametric tests or distribution-free test of hypotheses.

Parametric tests usually assume certain properties of the parent population from which we draw samples. Assumptions like observations come from a normal population, sample size is large, assumptions about the population parameters like mean, variance, etc., must hold good before parametric tests can be used. But there are situations when the researcher cannot or does not want to make such assumptions. In such situations we use statistical methods for testing hypotheses which are called non-parametric tests because such tests do not depend on any assumption about the parameters of the parent population. Besides, most non-parametric tests assume only nominal or ordinal data, whereas parametric tests require measurement equivalent to at least an interval scale. As a result, non-parametric tests need more observations than parametric tests to achieve the same size of Type I and Type II errors.[4] We take up in the present chapter some of the important parametric tests, whereas non-parametric tests will be dealt with in a separate chapter later in the book.

IMPORTANT PARAMETRIC TESTS

The important parametric tests are: (1) *z*-test; (2) *t*-test; (*3) χ^2-test, and (4) *F*-test. All these tests are based on the assumption of normality i.e., the source of data is considered to be normally distributed.

Mean of the population can be tested presuming different situations such as the population may be normal or other than normal, it may be finite or infinite, sample size may be large or small, variance of the population may be known or unknown and the alternative hypothesis may be two-sided or one-sided. Our testing technique will differ in different situations. We may consider some of the important situations.

1. *Population normal, population infinite, sample size may be large or small but variance of the population is known, H_a may be one-sided or two-sided*:

   In such a situation *z*-test is used for testing hypothesis of mean and the test statistic *z* is worked our as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

2. *Population normal, population finite, sample size may be large or small but variance of the population is known, H_a may be one-sided or two-sided*:

   In such a situation *z*-test is used and the test statistic *z* is worked out as under (using finite population multiplier):

$$z = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_p / \sqrt{n}\right) \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

and $$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{(n-1)}}$$

5. *Population may not be normal but sample size is large, variance of the population may be known or unknown, and H_a may be one-sided or two-sided:*

In such a situation we use z-test and work out the test statistic z as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}}$$

(This applies in case of infinite population when variance of the population is known but when variance is not known, we use σ_s in place of σ_p in this formula.)

OR

$$z = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_p/\sqrt{n}\right) \times \sqrt{(N-n)/(N-1)}}$$

(This applies in case of finite population when variance of the population is known but when variance is not known, we use σ_s in place of σ_p in this formula.)

3. *Population normal, population infinite, sample size small and variance of the population unknown, H_a may be one-sided or two-sided:*

In such a situation t-test is used and the test statistic t is worked out as under:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s/\sqrt{n}} \text{ with d.f.} = (n-1)$$

and $$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{(n-1)}}$$

4. *Population normal, population finite, sample size small and variance of the population unknown, and H_a may be one-sided or two-sided:*

In such a situation t-test is used and the test statistic 't' is worked out as under (using finite population multiplier):

$$t = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_s/\sqrt{n}\right) \times \sqrt{(N-n)/(N-1)}} \text{ with d.f.} = (n-1)$$

and $$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{(n-1)}}$$

5. *Population may not be normal but sample size is large, variance of the population may be known or unknown, and H_a may be one-sided or two-sided:*

In such a situation we use z-test and work out the test statistic z as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}}$$

(This applies in case of infinite population when variance of the population is known but when variance is not known, we use σ_s in place of σ_p in this formula.)

OR

$$z = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_p/\sqrt{n}\right) \times \sqrt{(N-n)/(N-1)}}$$

(This applies in case of finite population when variance of the population is known but when variance is not known, we use σ_s in place of σ_p in this formula.)

Example 1:

A sample of 400 male students is found to have a mean height 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation 1.30 inches? Test at 5% level of significance.

Taking the null hypothesis that the mean height of the population is equal to 67.39 inches, we can write:

$$H_0 : \mu_{H_0} = 67.39''$$
$$H_a : \mu_{H_0} \neq 67.39''$$

and the given information as $\bar{X} = 67.47''$, $\sigma_p = 1.30''$, $n = 400$. Assuming the population to be normal, we can work out the test statistic z as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}} = \frac{67.47 - 67.39}{1.30/\sqrt{400}} = \frac{0.08}{0.065} = 1.231$$

As H_a is two-sided in the given question, we shall be applying a two-tailed test for determining the rejection regions at 5% level of significance which comes to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is 1.231 which is in the acceptance region since $R : |z| > 1.96$ and thus H_0 is accepted. We may conclude that the given sample (with mean height = 67.47″) can be regarded to have been taken from a population with mean height 67.39″ and standard deviation 1.30″ at 5% level of significance.

Example 2:

Suppose we are interested in a population of 20 industrial units of the same size, all of which are experiencing excessive labour turnover problems. The past records show that the mean of the distribution of annual turnover is 320 employees, with a standard deviation of 75 employees. A sample of 5 of these industrial units is taken at random which gives a mean of annual turnover as 300 employees. Is the sample mean consistent with the population mean? Test at 5% level.

$$H_0: \mu_{H_0} = 320 \text{ employees}$$
$$H_a: \mu_{H_0} \neq 320 \text{ employees}$$

and the given information as under:

$$\bar{X} = 300 \text{ employees}, \ \sigma_p = 75 \text{ employees}$$
$$n = 5; N = 20$$

Assuming the population to be normal, we can work out the test statistic z as under:

$$z^* = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n} \times \sqrt{(N-n)/(N-1)}}$$

$$= \frac{300 - 320}{75/\sqrt{5} \times \sqrt{(20-5)/(20-1)}} = -\frac{20}{(33.54)\,(.888)}$$

$$= -0.67$$

As H_a is two-sided in the given question, we shall apply a two-tailed test for determining the rejection regions at 5% level of significance which comes to as under, using normal curve area table:

$$R: |z| > 1.96$$

The observed value of z is –0.67 which is in the acceptance region since $R: |z| > 1.96$ and thus, H_0 is accepted and we may conclude that the sample mean is consistent with population mean i.e., the population mean 320 is supported by sample results.

Example 3:

The mean of a certain production process is known to be 50 with a standard deviation of 2.5. The production manager may welcome any change is mean value towards higher side but would like to safeguard against decreasing values of mean. He takes a sample of 12 items that gives a mean value of 48.5. What inference should the manager take for the production process on the basis of sample results? Use 5 per cent level of significance for the purpose.

$$H_0: \mu_{H_0} = 50$$

$$H_a: \mu_{H_0} < 50 \text{ (Since the manager wants to safeguard against decreasing values of mean.)}$$

and the given information as $\bar{X} = 48.5$, $\sigma_p = 2.5$ and $n = 12$. Assuming the population to be normal, we can work out the test statistic z as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}} = \frac{48.5 - 50}{2.5/\sqrt{12}} = -\frac{1.5}{(2.5)/(3.464)} = -2.0784$$

As H_a is one-sided in the given question, we shall determine the rejection region applying one-tailed test (in the left tail because H_a is of less than type) at 5 per cent level of significance and it comes to as under, using normal curve area table:

$$R: z < -1.645$$

The observed value of z is – 2.0784 which is in the rejection region and thus, H_0 is rejected at 5 per cent level of significance. We can conclude that the production process is showing mean which is significantly less than the population mean and this calls for some corrective action concerning the said process.

Example 4:

Raju Restaurant near the railway station at Falna has been having average sales of 500 tea cups per day. Because of the development of bus stand nearby, it expects to increase its sales. During the first 12 days after the start of the bus stand, the daily sales were as under:

550, 570, 490, 615, 505, 580, 570, 460, 600, 580, 530, 526

On the basis of this sample information, can one conclude that Raju Restaurant's sales have increased?

Use 5 per cent level of significance.

H_0 : μ = 500 cups per day

H_a : μ > 500 (as we want to conclude that sales have increased).

As the sample size is small and the population standard deviation is not known, we shall use *t*-test assuming normal population and shall work out the test statistic *t* as:

$$t = \frac{\bar{X} - \mu}{\sigma_s/\sqrt{n}}$$

(To find $\bar{X}$ and σ_s we make the following computations:)

$$\therefore \quad \bar{X} = \frac{\Sigma X_i}{n} = \frac{6576}{12} = 548$$

and

$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

Hence,

$$t = \frac{548 - 500}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$$

Degree of freedom = $n - 1 = 12 - 1 = 11$

As H_a is one-sided, we shall determine the rejection region applying one-tailed test (in the right tail because H_a is of more than type) at 5 per cent level of significance and it comes to as under, using table of *t*-distribution for 11 degrees of freedom:

$$R : t > 1.796$$

The observed value of *t* is 3.558 which is in the rejection region and thus H_0 is rejected at 5 per cent level of significance and we can conclude that the sample data indicate that Raju restaurant's sales have increased.

HYPOTHESIS TESTING FOR DIFFERENCES BETWEEN MEANS

In many decision-situations, we may be interested in knowing whether the parameters of two populations are alike or different. For instance, we may be interested in testing whether female workers earn less than male workers for the same job. We shall explain now the technique of

hypothesis testing for differences between means. The null hypothesis for testing of difference between means is generally stated as $H_0 : \mu_1 = \mu_2$, where μ_1 is population mean of one population and μ_2 is population mean of the second population, assuming both the populations to be normal populations. Alternative hypothesis may be of not equal to or less than or greater than type as stated earlier and accordingly we shall determine the acceptance or rejection regions for testing the hypotheses. There may be different situations when we are examining the significance of difference between two means, but the following may be taken as the usual situations:

1. *Population variances are known or the samples happen to be large samples:*
   In this situation we use z-test for difference in means and work out the test statistic z as under:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_{p1}^2}{n_1} + \frac{\sigma_{p2}^2}{n_2}}}$$

   In case σ_{p_1} and σ_{p_2} are not known, we use σ_{s_1} and σ_{s_2} respectively in their places calculating

$$\sigma_{s_1} = \sqrt{\frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1}} \text{ and } \sigma_{s_2} = \sqrt{\frac{\Sigma\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1}}$$

2. *Samples happen to be large but presumed to have been drawn from the same population whose variance is known:*
   In this situation we use z test for difference in means and work out the test statistic z as under:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

   In case σ_p is not known, we use $\sigma_{s_{1.2}}$ (combined standard deviation of the two samples) in its place calculating

$$\sigma_{s_{1.2}} = \sqrt{\frac{n_1\left(\sigma_{s_1}^2 + D_1^2\right) + n_2\left(\sigma_{s_2}^2 + D_2^2\right)}{n_1 + n_2}}$$

   where $D_1 = \left(\bar{X}_1 - \bar{X}_{1.2}\right)$

   $D_2 = \left(\bar{X}_2 - \bar{X}_{1.2}\right)$

$$\bar{X}_{1.2} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

3. *Samples happen to be small samples and population variances not known but assumed to be equal:*

In this situation we use *t*-test for difference in means and work out the test statistic *t* as under:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with d.f. = $(n_1 + n_2 - 2)$

Alternatively, we can also state

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)\sigma_{s_1}^2 + (n_2 - 1)\sigma_{s_2}^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with d.f. = $(n_1 + n_2 - 2)$

Example 1:

The mean produce of wheat of a sample of 100 fields in 200 lbs. per acre with a standard deviation of 10 lbs. Another samples of 150 fields gives the mean of 220 lbs. with a standard deviation of 12 lbs. Can the two samples be considered to have been taken from the same population whose standard deviation is 11 lbs? Use 5 per cent level of significance.

$$H_0 : \mu = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

and the given information as $n_1 = 100$; $n_2 = 150$;

$$\bar{X}_1 = 200 \text{ lbs.}; \qquad \bar{X}_2 = 220 \text{ lbs.};$$

$$\sigma_{s_1} = 10 \text{ lbs.}; \qquad \sigma_{s_2} = 12 \text{ lbs.};$$

and $$\sigma_p = 11 \text{ lbs.}$$

Assuming the population to be normal, we can work out the test statistic *z* as under:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{200 - 220}{\sqrt{(11)^2\left(\frac{1}{100} + \frac{1}{150}\right)}}$$

$$= -\frac{20}{1.42} = -14.08$$

As H_a is two-sided, we shall apply a two-tailed test for determining the rejection regions at 5 per cent level of significance which come to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of *z* is – 14.08 which falls in the rejection region and thus we reject H_0 and conclude that the two samples cannot be considered to have been taken at 5 per cent level of significance from the same population whose standard deviation is 11 lbs. This means that the difference between means of two samples is statistically significant and not due to sampling fluctuations.

Example 2:

A group of seven-week old chickens reared on a high protein diet weigh 12, 15, 11, 16, 14, 14, and 16 ounces; a second group of five chickens, similarly treated except that they receive a low protein diet, weigh 8, 10, 14, 10 and 13 ounces. Test at 5 per cent level whether there is significant evidence that additional protein has increased the weight of the chickens. Use assumed mean (or A_1) = 10 for the sample of 7 and assumed mean (or A_2) = 8 for the sample of 5 chickens in your calculations.

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 > \mu_2$ (as we want to conclude that additional protein has increased the weight of chickens)

Since in the given question variances of the populations are not known and the size of samples is small, we shall use *t*-test for difference in means, assuming the populations to be normal and thus work out the test statistic *t* as under:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)\sigma_{s_1}^2 + (n_2 - 1)\sigma_{s_2}^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with d.f. = $(n_1 + n_2 - 2)$

From the sample data we work out $\bar{X}_1$, $\bar{X}_2$, $\sigma_{s_1}^2$ and $\sigma_{s_2}^2$ (taking high protein diet sample as sample one and low protein diet sample as sample two) as shown below:

$$\therefore \quad \bar{X}_1 = A_1 + \frac{\Sigma(X_{1i} - A_1)}{n_1} = 10 + \frac{28}{7} = 14 \text{ ounces}$$

$$\bar{X}_2 = A_2 + \frac{\Sigma(X_{2i} - A_2)}{n_2} = 8 + \frac{15}{5} = 11 \text{ ounces}$$

$$\sigma_{s_1}^2 = \frac{\Sigma(X_{1i} - A_1)^2 - \left[\Sigma(X_{1i} - A_1)\right]^2 / n_1}{(n_1 - 1)}$$

$$= \frac{134 - (28)^2/7}{7 - 1} = 3.667 \text{ ounces}$$

$$\sigma_{s_2}^2 = \frac{\Sigma(X_{2i} - A_2)^2 - \left[\Sigma(X_{2i} - A_2)\right]^2 / n_2}{(n_2 - 1)}$$

$$= \frac{69 - (15)^2/5}{5 - 1} = 6 \text{ ounces}$$

Hence,
$$t = \frac{14 - 11}{\sqrt{\frac{(7-1)(3.667) + (5-1)(6)}{7 + 5 - 2}} \times \sqrt{\frac{1}{7} + \frac{1}{5}}}$$

$$= \frac{3}{\sqrt{4.6} \times \sqrt{.345}} = \frac{3}{1.26} = 2.381$$

Degrees of freedom = $(n_1 + n_2 - 2) = 10$

As H_a is one-sided, we shall apply a one-tailed test (in the right tail because H_a is of more than type) for determining the rejection region at 5 per cent level which comes to as under, using table of t-distribution for 10 degrees of freedom:

$$R : t > 1.812$$

The observed value of t is 2.381 which falls in the rejection region and thus, we reject H_0 and conclude that additional protein has increased the weight of chickens, at 5 per cent level of significance.

HYPOTHESIS TESTING FOR COMPARING TWO RELATED SAMPLES

Paired *t*-test is a way to test for comparing two related samples, involving small values of *n* that does not require the variances of the two populations to be equal, but the assumption that the two populations are normal must continue to apply. For a paired *t*-test, it is necessary that the observations in the two samples be collected in the form of what is called matched pairs i.e., "each observation in the one sample must be paired with an observation in the other sample in such a manner that these observations are somehow "matched" or related, in an attempt to eliminate extraneous factors which are not of interest in test."[5] Such a test is generally considered appropriate in a before-and-after-treatment study. For instance, we may test a group of certain students before and after training in order to know whether the training is effective, in which situation we may use paired *t*-test. To apply this test, we first work out the difference score for each matched pair, and then find out the average of such differences, $\overline{D}$, along with the sample variance of the difference score. If the values from the two matched samples are denoted as X_i and Y_i and the differences by $D_i (D_i = X_i - Y_i)$, then the mean of the differences i.e.,

$$\overline{D} = \frac{\Sigma D_i}{n}$$

and the variance of the differences or

$$\left(\sigma_{diff.}\right)^2 = \frac{\Sigma D_i^2 - \left(\overline{D}\right)^2 \cdot n}{n-1}$$

Assuming the said differences to be normally distributed and independent, we can apply the paired *t*-test for judging the significance of mean of differences and work out the test statistic *t* as under:

$$t = \frac{\overline{D} - 0}{\sigma_{diff}/\sqrt{n}} \quad \text{with } (n-1) \text{ degrees of freedom}$$

where $\overline{D}$ = Mean of differences

$\sigma_{diff.}$ = Standard deviation of differences

n = Number of matched pairs

Example1 :

Memory capacity of 9 students was tested before and after training. State at 5 per cent level of significance whether the training was effective from the following scores:

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Before	10	15	9	3	7	12	16	17	4
After	12	17	8	5	6	11	18	20	3

Use paired *t*-test as well as *A*-test for your answer.

.

$H_0: \mu_1 = \mu_2$ which is equivalent to test $H_0: \overline{D} = 0$

$H_a: \mu_1 < \mu_2$ (as we want to conclude that training has been effective)

As we are having matched pairs, we use paired t-test and work out the test statistic t as under:

$$t = \frac{\overline{D} - 0}{\sigma_{diff.}/\sqrt{n}}$$

To find the value of t, we shall first have to work out the mean and standard deviation of differences as shown below:

Student	*Score before training* X_i	*Score after training* Y_i	*Difference* $(D_i = X_i - Y_i)$	*Difference Squared* D_i^2
1	10	12	−2	4
2	15	17	−2	4
3	9	8	1	1
4	3	5	−2	4
5	7	6	1	1
6	12	11	1	1
7	16	18	−2	4
8	17	20	−3	9
9	4	3	1	1
$n = 9$			$\Sigma D_i = -7$	$\Sigma D_i^2 = 29$

$\therefore$ Mean of Differences or $\overline{D} = \frac{\Sigma D_i}{n} = \frac{-7}{9} = -0.778$

and Standard deviation of differences or

$$\sigma_{diff.} = \sqrt{\frac{\Sigma D_i^2 - \left(\overline{D}\right)^2 \cdot n}{n-1}}$$

$$= \sqrt{\frac{29 - (-.778)^2 \times 9}{9-1}}$$

$$= \sqrt{2.944} = 1.715$$

Hence, $$t = \frac{-0.778 - 0}{1.715/\sqrt{9}} = \frac{-.778}{0.572} = -1.361$$

Degrees of freedom $= n - 1 = 9 - 1 = 8$.

As H_a is one-sided, we shall apply a one-tailed test (in the left tail because H_a is of less than type) for determining the rejection region at 5 per cent level which comes to as under, using the table of t-distribution for 8 degrees of freedom:

$$R: t < -1.860$$

The observed value of t is –1.361 which is in the acceptance region and thus, we accept H_0 and conclude that the difference in score before and after training is insignificant i.e., it is only due to sampling fluctuations. Hence we can infer that the training was not effective.

Example 2:

The sales data of an item in six shops before and after a special promotional campaign are:

Shops	*A*	*B*	*C*	*D*	*E*	*F*
Before the promotional campaign	53	28	31	48	50	42
After the campaign	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5 per cent level of significance. Use paired *t*-test as well as *A*-test.

$H_0: \mu_1 = \mu_2$ which is equivalent to test $H_0: \overline{D} = 0$

$H_a: \mu_1 < \mu_2$ (as we want to conclude that campaign has been a success).

Because of the matched pairs we use paired *t*-test and work out the test statistic '*t*' as under:

$$t = \frac{\overline{D} - 0}{\sigma_{diff}/\sqrt{n}}$$

To find the value of *t*, we first work out the mean and standard deviation of differences as under:

Shops	*Sales before campaign* X_i	*Sales after campaign* Y_i	*Difference* $(D_i = X_i - Y_i)$	*Difference squared* D_i^2
A	53	58	–5	25
B	28	29	–1	1
C	31	30	1	1
D	48	55	–7	49
E	50	56	–6	36
F	42	45	–3	9
$n = 6$			$\Sigma D_i = -21$	$\Sigma D_i^2 = 121$

$$\therefore \quad \overline{D} = \frac{\Sigma D_i}{n} = -\frac{21}{6} = -3.5$$

$$\sigma_{diff.} = \sqrt{\frac{\Sigma D_i^2 - \left(\overline{D}\right)^2 \cdot n}{n-1}} = \sqrt{\frac{121 - (-3.5)^2 \times 6}{6-1}} = 3.08$$

Hence, $$t = \frac{-3.5 - 0}{3.08/\sqrt{6}} = \frac{-3.5}{1.257} = -2.784$$

Degrees of freedom = $(n - 1) = 6 - 1 = 5$

As H_a is one-sided, we shall apply a one-tailed test (in the left tail because H_a is of less than type) for determining the rejection region at 5 per cent level of significance which come to as under, using table of t-distribution for 5 degrees of freedom:

$$R : t < -2.015$$

The observed value of t is – 2.784 which falls in the rejection region and thus, we reject H_0 at 5 per cent level and conclude that sales promotional campaign has been a success.

HYPOTHESIS TESTING OF PROPORTIONS

In case of qualitative phenomena, we have data on the basis of presence or absence of an attribute(s). With such data the sampling distribution may take the form of binomial probability distribution whose mean would be equal to $n \cdot p$ and standard deviation equal to $\sqrt{n \cdot p \cdot q}$, where p represents the probability of success, q represents the probability of failure such that $p + q = 1$ and n, the size of the sample. Instead of taking mean number of successes and standard deviation of the number of successes, we may record the proportion of successes in each sample in which case the mean and standard deviation (or the standard error) of the sampling distribution may be obtained as follows:

Mean proportion of successes $= (n \cdot p)/n = p$

and standard deviation of the proportion of successes $= \sqrt{\frac{p \cdot q}{n}}$.

In n is large, the binomial distribution tends to become normal distribution, and as such for proportion testing purposes we make use of the test statistic z as under:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where $\hat{p}$ is the sample proportion.

For testing of proportion, we formulate H_0 and H_a and construct rejection region, presuming normal approximation of the binomial distribution, for a predetermined level of significance and then may judge the significance of the observed sample result. The following examples make all this quite clear.

Example 1:

A sample survey indicates that out of 3232 births, 1705 were boys and the rest were girls. Do these figures confirm the hypothesis that the sex ratio is 50 : 50? Test at 5 per cent level of significance.

$$H_0 : p = p_{H_0} = \frac{1}{2}$$

$$H_a : p \neq p_{H_0}$$

Hence the probability of boy birth or $p = \frac{1}{2}$ and the probability of girl birth is also $\frac{1}{2}$.

Considering boy birth as success and the girl birth as failure, we can write as under:

the proportion success or $p = \frac{1}{2}$

the proportion of failure or $q = \frac{1}{2}$

and $n = 3232$ (given).

The standard error of proportion of success.

$$= \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{3232}} = 0.0088$$

Observed sample proportion of success, or

$$\hat{p} = 1705/3232 = 0.5275$$

and the test statistic

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.5275 - .5000}{.0088} = 3.125$$

As H_a is two-sided in the given question, we shall be applying the two-tailed test for determining the rejection regions at 5 per cent level which come to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is 3.125 which comes in the rejection region since $R : |z| > 1.96$ and thus, H_0 is rejected in favour of H_a. Accordingly, we conclude that the given figures do not conform the hypothesis of sex ratio being 50 : 50.

Example 2:

The null hypothesis is that 20 per cent of the passengers go in first class, but management recognizes the possibility that this percentage could be more or less. A random sample of 400 passengers includes 70 passengers holding first class tickets. Can the null hypothesis be rejected at 10 per cent level of significance?

$H_0 : p = 20\%$ or 0.20

and $H_a : p \neq 20\%$

H

Hence, $p = 0.20$ and
$q = 0.80$

Observed sample proportion $(\hat{p}) = 70/400 = 0.175$

and the test statistic $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p \cdot q}{n}}} = \dfrac{0.175 - .20}{\sqrt{\dfrac{.20 \times .80}{400}}} = -1.25$

As H_a is two-sided we shall determine the rejection regions applying two-tailed test at 10 per cent level which come to as under, using normal curve area table:

$$R : |z| > 1.645$$

The observed value of z is –1.25 which is in the acceptance region and as such H_0 is accepted. Thus the null hypothesis cannot be rejected at 10 per cent level of significance.

HYPOTHESIS TESTING FOR DIFFERENCE BETWEEN PROPORTIONS

If two samples are drawn from different populations, one may be interested in knowing whether the difference between the proportion of successes is significant or not. In such a case, we start with the hypothesis that the difference between the proportion of success in sample one $(\hat{p}_1)$ and the proportion of success in sample two $(\hat{p}_2)$ is due to fluctuations of random sampling. In other words, we take the null hypothesis as $H_0 : \hat{p}_1 = \hat{p}_2$ and for testing the significance of difference, we work out the test statistic as under:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \dfrac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}}$$

where $\hat{p}_1$ = proportion of success in sample one

$\hat{p}_2$ = proportion of success in sample two

$\hat{q}_1 = 1 - \hat{p}_1$

$\hat{q}_2 = 1 - \hat{p}_2$

n_1 = size of sample one

n_2 = size of sample two

and

$\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}$ = the standard error of difference between two sample proportions.*

Example1:

A drug research experimental unit is testing two drugs newly developed to reduce blood pressure levels. The drugs are administered to two different sets of animals. In group one, 350 of 600 animals tested respond to drug one and in group two, 260 of 500 animals tested respond to drug two. The research unit wants to test whether there is a difference between the efficacy of the said two drugs at 5 per cent level of significance. How will you deal with this problem?

$H_0: \hat{p}_1 = \hat{p}_2$

The alternative hypothesis can be taken as that there is a difference between the drugs i.e., $H_a: \hat{p}_1 \neq \hat{p}_2$ and the given information can be stated as:

$$\hat{p}_1 = 350/600 = 0.583$$

$$\hat{q}_1 = 1 - \hat{p}_1 = 0.417$$

$$n_1 = 600$$

$$\hat{p}_2 = 260/500 = 0.520$$

$$\hat{q}_2 = 1 - \hat{p}_2 = 0.480$$

$$n_2 = 500$$

We can work out the test statistic z thus:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} = \frac{0.583 - 0.520}{\sqrt{\frac{(.583)(.417)}{600} + \frac{(.520)(.480)}{500}}}$$

$$= 2.093$$

As H_a is two-sided, we shall determine the rejection regions applying two-tailed test at 5% level which comes as under using normal curve area table:

$$R: |z| > 1.96$$

The observed value of z is 2.093 which is in the rejection region and thus, H_0 is rejected in favour of H_a and as such we conclude that the difference between the efficacy of the two drugs is significant.

Example 2:

At a certain date in a large city 400 out of a random sample of 500 men were found to be smokers. After the tax on tobacco had been heavily increased, another random sample of 600 men in the same city included 400 smokers. Was the observed decrease in the proportion of smokers significant? Test at 5 per cent level of significance.

Solution:

on tobacco remains unchanged i.e. $H_0: \hat{p}_1 = \hat{p}_2$ and the alternative hypothesis that proportion of smokers after tax has decreased i.e.,

$$H_a: \hat{p}_1 > \hat{p}_2$$

On the presumption that the given populations are similar as regards the given attribute, we work out the best estimate of proportion of smokers (p_0) in the population as under, using the given information:

$$p_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{500\left(\frac{400}{500}\right) + 600\left(\frac{400}{600}\right)}{500 + 600} = \frac{800}{1100} = \frac{8}{11} = .7273$$

Thus, $q_0 = 1 - p_0 = .2727$

The test statistic z can be worked out as under:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_0q_0}{n_1} + \frac{p_0q_0}{n_2}}} = \frac{\frac{400}{500} - \frac{400}{600}}{\sqrt{\frac{(.7273)(.2727)}{500} + \frac{(.7273)(.2727)}{600}}}$$

$$= \frac{0.133}{0.027} = 4.926$$

As the H_a is one-sided we shall determine the rejection region applying one-tailed test (in the right tail because H_a is of greater than type) at 5 per cent level and the same works out to as under, using normal curve area table:

$$R: z > 1.645$$

The observed value of z is 4.926 which is in the rejection region and so we reject H_0 in favour of H_a and conclude that the proportion of smokers after tax has decreased significantly.

HYPOTHESIS TESTING FOR COMPARING A VARIANCE TO SOME HYPOTHESISED POPULATION VARIANCE

The test we use for comparing a sample variance to some theoretical or hypothesised variance of population is different than *z*-test or the *t*-test. The test we use for this purpose is known as chi-square test and the test statistic symbolised as χ^2, known as the chi-square value, is worked out. The chi-square value to test the null hypothesis viz, $H_0: \sigma_s^2 = \sigma_p^2$ worked out as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

where σ_s^2 = variance of the sample

σ_p^2 = variance of the population

$(n-1)$ = degree of freedom, n being the number of items in the sample.

Then by comparing the calculated value of χ^2 with its table value for $(n-1)$ degrees of freedom at a given level of significance, we may either accept H_0 or reject it. If the calculated value of χ^2 is equal to or less than the table value, the null hypothesis is accepted; otherwise the null hypothesis is rejected. This test is based on chi-square distribution which is not symmetrical and all the values happen to be positive; one must simply know the degrees of freedom for using such a distribution.*

TESTING THE EQUALITY OF VARIANCES OF TWO NORMAL POPULATIONS

When we want to test the equality of variances of two normal populations, we make use of *F*-test based on *F*-distribution. In such a situation, the null hypothesis happens to be $H_0: \sigma_{p_1}^2 = \sigma_{p_2}^2$, $\sigma_{p_1}^2$ and $\sigma_{p_2}^2$ representing the variances of two normal populations. This hypothesis is tested on the basis of sample data and the test statistic F is found, using $\sigma_{s_1}^2$ and $\sigma_{s_2}^2$ the sample estimates for $\sigma_{p_1}^2$ and $\sigma_{p_2}^2$ respectively, as stated below:

$$F = \frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$$

where $\sigma_{s_1}^2 = \dfrac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2}{(n_1 - 1)}$ and $\sigma_{s_2}^2 = \dfrac{\Sigma\left(X_{2i} - \bar{X}_2\right)^2}{(n_2 - 1)}$

While calculating F, $\sigma^2_{s_1}$ is treated $> \sigma^2_{s_2}$ which means that the numerator is always the greater variance. Tables for F-distribution** have been prepared by statisticians for different values of F at different levels of significance for different degrees of freedom for the greater and the smaller variances. By comparing the observed value of F with the corresponding table value, we can infer whether the difference between the variances of samples could have arisen due to sampling fluctuations. If the calculated value of F is greater than table value of F at a certain level of significance for $(n_1 - 1)$ and $(n_2 - 2)$ degrees of freedom, we regard the F-ratio as significant. Degrees of freedom for greater variance is represented as v_1 and for smaller variance as v_2. On the other hand, if the calculated value of F is smaller than its table value, we conclude that F-ratio is not significant. If F-ratio is considered non-significant, we accept the null hypothesis, but if F-ratio is considered significant, we then reject H_0 (i.e., we accept H_a).

When we use the F-test, we presume that

(i) the populations are normal;
(ii) samples have been drawn randomly;
(iii) observations are independent; and
(iv) there is no measurement error.

Example 1:

Two random samples drawn from two normal populations are:

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

Test using variance ratio at 5 per cent and 1 per cent level of significance whether the two populations have the same variances.

Solution:

drawn have the same variances i.e., $H_0 : \sigma^2_{p_1} = \sigma^2_{p_2}$. From the sample data we work out $\sigma^2_{s_1}$ and $\sigma^2_{s_2}$ as under:

Sample 1			*Sample 2*		
X_{1i}	$(X_{1i} - \bar{X}_1)$	$(X_{1i} - \bar{X}_1)^2$	X_{2i}	$(X_{2i} - \bar{X}_2)$	$(X_{2i} - \bar{X}_2)^2$
20	–2	4	27	–8	64
16	–6	36	33	–2	4
26	4	16	42	7	49
27	5	25	35	0	0
23	1	1	32	–3	9
22	0	0	34	–1	1
18	–4	16	38	3	9
24	2	4	28	–7	49
25	3	9	41	6	36
19	–3	9	43	8	64
			30	–5	25
			37	2	4
$\Sigma X_{1i} = 220$		$\Sigma(X_{1i} - \bar{X}_1)^2 = 120$	$\Sigma X_{2i} = 420$		$\Sigma(X_{2i} - \bar{X}_2)^2 = 314$
$n_1 = 10$			$n_2 = 12$		

$$\bar{X}_1 = \frac{\Sigma X_{1i}}{n_1} = \frac{220}{10} = 22; \qquad \bar{X}_2 = \frac{\Sigma X_{2i}}{n_2} = \frac{420}{12} = 35$$

$$\therefore \qquad \sigma_{s_1}^2 = \frac{\Sigma(X_{1i} - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{10 - 1} = 13.33$$

and
$$\sigma_{s_2}^2 = \frac{\Sigma(X_{2i} - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{12 - 1} = 28.55$$

Hence,
$$F = \frac{\sigma_{s_2}^2}{\sigma_{s_1}^2} \qquad \left(\because \sigma_{s_2}^2 > \sigma_{s_1}^2\right)$$

$$= \frac{28.55}{13.33} = 2.14$$

Degrees of freedom in sample 1 = $(n_1 - 1) = 10 - 1 = 9$

Degrees of freedom in sample 2 = $(n_2 - 1) = 12 - 1 = 11$

As the variance of sample 2 is greater variance, hence

$$v_1 = 11;\ v_2 = 9$$

The table value of F at 5 per cent level of significance for $v_1 = 11$ and $v_2 = 9$ is 3.11 and the table value of F at 1 per cent level of significance for $v_1 = 11$ and $v_2 = 9$ is 5.20.

Since the calculated value of $F = 2.14$ which is less than 3.11 and also less than 5.20, the F ratio is insignificant at 5 per cent as well as at 1 per cent level of significance and as such we accept the null hypothesis and conclude that samples have been drawn from two populations having the same variances.

Example 2:

Given $n_1 = 9$; $n_2 = 8$

$$\Sigma\left(X_{1i} - \bar{X}_1\right)^2 = 184$$

$$\Sigma\left(X_{2i} - \bar{X}_2\right)^2 = 38$$

Apply *F*-test to judge whether this difference is significant at 5 per cent level.

To test this, we work out the *F*-ratio as under:

$$F = \frac{\sigma^2_{s_1}}{\sigma^2_{s_2}} = \frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2 / (n_1 - 1)}{\Sigma\left(X_{2i} - \bar{X}_2\right)^2 / (n_2 - 1)}$$

$$= \frac{184/8}{38/7} = \frac{23}{5.43} = 4.25$$

$v_1 = 8$ being the number of d.f. for greater variance
$v_2 = 7$ being the number of d.f. for smaller variance.

The table value of *F* at 5 per cent level for $v_1 = 8$ and $v_2 = 7$ is 3.73. Since the calculated value of *F* is greater than the table value, the *F* ratio is significant at 5 per cent level. Accordingly we reject H_0 and conclude that the difference is significant.

QUESTIONS:

1. A coin is tossed 10,000 times and head turns up 5,195 times. Is the coin unbiased?
2. In some dice throwing experiments, *A* threw dice 41952 times and of these 25145 yielded a 4 or 5 or 6. Is this consistent with the hypothesis that the dice were unbiased?
3. A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out three imperfect articles in a batch of 100. Has the machine improved? Test at 5% level of significance.
4. In two large populations, there are 35% and 30% respectively fair haired people. Is this difference likely to be revealed by simple sample of 1500 and 1000 respectively from the two populations?
5. In a certain association table the following frequencies were obtained: (*AB*) = 309, (*Ab*) = 214, (*aB*) = 132, (*ab*) = 119.
   Can the association between *AB* as per the above data can be said to have arisen as a fluctuation of simple sampling?
6. A sample of 900 members is found to have a mean of 3.47 cm. Can it be reasonably regarded as a simple sample from a large population with mean 3.23 cm. and standard deviation 2.31 cm.?

7. The means of the two random samples of 1000 and 2000 are 67.5 and 68.0 inches respectively. Can the samples be regarded to have been drawn from the same population of standard deviation 9.5 inches? Test at 5% level of significance.

8. A large corporation uses thousands of light bulbs every year. The brand that has been used in the past has an average life of 1000 hours with a standard deviation of 100 hours. A new brand is offered to the corporation at a price far lower than one they are paying for the old brand. It is decided that they will switch to the new brand unless it is proved with a level of significance of 5% that the new brand has smaller average life than the old brand. A random sample of 100 new brand bulbs is tested yielding an observed sample mean of 985 hours. Assuming that the standard deviation of the new brand is the same as that of the old brand,

(a) What conclusion should be drawn and what decision should be made?

(b) What is the probability of accepting the new brand if it has the mean life of 950 hours?

9. Ten students are selected at random from a school and their heights are found to be, in inches, 50, 52, 52, 53, 55, 56, 57, 58, 58 and 59. In the light of these data, discuss the suggestion that the mean height of the students of the school is 54 inches. You may use 5% level of significance (Apply *t*-test as well as *A*-test).

10. In a test given to two groups of students, the marks obtained were 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 difference between mean marks obtained by students of the above two groups. Test at five per cent level of significance.

11. The heights of six randomly chosen sailors are, in inches, 63, 65, 58, 69, 71 and 72. The heights of 10 randomly chosen soldiers are, in inches, 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Do these figures indicate that soldiers are on an average shorter than sailors? Test at 5% level of significance.

12. Ten young recruits were put through a strenuous physical training programme by the army. Their weights (in kg) were recorded before and after with the following results:

Recruit	1	2	3	4	5	6	7	8	9	10
Weight before	127	195	162	170	143	205	168	175	197	136
Weight after	135	200	160	182	147	200	172	186	194	141

Using 5% level of significance, should we conclude that the programme affects the average weight of young recruits (Answer using *t*-test as well as *A*-test

13. Answer using *F*-test whether the following two samples have come from the same population:

Sample 1 17 27 18 25 27 29 27 23 17

Sample 2 16 16 20 16 20 17 15 21

Use 5% level of significance.

14. The following table gives the number of units produced per day by two workers *A* and *B* for a number of days:

A 40 30 38 41 38 35

B 39 38 41 33 32 49 49 34

Should these results be accepted as evidence that *B* is the more stable worker? Use *F*-test at 5% level.

15. A sample of 600 persons selected at random from a large city gives the result that males are 53%. Is there reason to doubt the hypothesis that males and females are in equal numbers in the city? Use 1% level of significance.

16. 12 students were given intensive coaching and 5 tests were conducted in a month. The scores of tests 1 and 5 are given below. Does the score from Test 1 to Test 5 show an improvement? Use 5% level of significance.

No. of students	1	2	3	4	5	6	7	8	9	10	11	12
Marks in 1st Test	50	42	51	26	35	42	60	41	70	55	62	38
Marks in 5th test	62	40	61	35	30	52	68	51	84	63	72	50

Chi-Square Test

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Chi-square, symbolically written as χ^2 (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance. As a non-parametric* test, it "can be used to determine if categorical data shows dependency or the two classifications are independent. It can also be used to make comparisons between theoretical populations and actual data when categories are used."[1] Thus, the chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for all researchers to (i) test the goodness of fit; (ii) test the significance of association between two attributes, and (iii) test the homogeneity or the significance of population variance.

CHI-SQUARE AS A TEST FOR COMPARING VARIANCE

The chi-square value is often used to judge the significance of population variance i.e., we can use the test to judge if a random sample has been drawn from a normal population with mean (μ) and with a specified variance (σ_p^2). The test is based on χ^2-distribution. Such a distribution we encounter when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and, thus, have distributions that are related to χ^2-distribution. If we take each one of a collection of sample variances, divided them by the known population variance and multiply these quotients by $(n-1)$, where n means the number of items in the sample, we shall obtain a χ^2-distribution. Thus, $\frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\sigma_s^2}{\sigma_p^2}$ (d.f.) would have the same distribution as χ^2-distribution with $(n-1)$ degrees of freedom.

Example 1:

Weight of 10 students is as follows:

S. No.	1	2	3	4	5	6	7	8	9	10
Weight (kg.)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the above sample of 10 students was drawn is equal to 20 kgs? Test this at 5 per cent and 1 per cent level of significance.

S. No.	X_i (Weight in kgs.)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	38	–9	81
2	40	–7	49
3	45	–2	04
4	53	+6	36
5	47	+0	00
6	43	–4	16
7	55	+8	64
8	48	+1	01
9	52	+5	25
10	49	+2	04
$n = 10$	$\Sigma X_i = 470$	$\Sigma(X_i - \bar{X})^2 = 280$	

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{470}{10} = 47 \text{ kgs.}$$

$$\therefore \quad \sigma_s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{280}{10-1}} = \sqrt{31.11}$$

or $$\sigma_s^2 = 31.11.$$

Let the null hypothesis be $H_0 : \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

$$= \frac{31.11}{20}(10-1) = 13.999.$$

Degrees of freedom in the given case is $(n-1) = (10-1) = 9$. At 5 per cent level of significance the table value of $\chi^2 = 16.92$ and at 1 per cent level of significance, it is 21.67 for 9 d.f. and both these values are greater than the calculated value of χ^2 which is 13.999. Hence we accept the null hypothesis and conclude that the variance of the given distribution can be taken as 20 kgs at 5 per cent as also at 1 per cent level of significance. In other words, the sample can be said to have been taken from a population with variance 20 kgs.

Example 2:

A sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5 at 5 per cent level of significance.

$$n = 10$$

$$\Sigma\left(X_i - \bar{X}\right)^2 = 50$$

$$\therefore \qquad \sigma_s^2 = \frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1} = \frac{50}{9}$$

Take the null hypothesis as $H_0: \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis, we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\frac{50}{9}}{5}(10-1) = \frac{50}{9} \times \frac{1}{5} \times \frac{9}{1} = 10$$

Degrees of freedom = (10 – 1) = 9.

The table value of χ^2 at 5 per cent level for 9 d.f. is 16.92. The calculated value of χ^2 is less than this table value, so we accept the null hypothesis and conclude that the variance of the population is 5 as given in the question.

CHI-SQUARE AS A NON-PARAMETRIC TEST

Chi-square is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom (implicitly of course the size of the sample) for using this test. As a non-parametric test, chi-square can be used (i) as a test of goodness of fit and (ii) as a test of independence.

As a test of goodness of fit, χ^2 test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data. When some theoretical distribution is fitted to the given data, we are always interested in knowing as to how well this distribution fits with the observed data. The chi-square test can give answer to this. If the calculated value of χ^2 is less than the table value at a certain level of significance, the fit is considered to be a good one which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the calculated value of χ^2 is greater than its table value, the fit is not considered to be a good one.

As a test of independence, χ^2 test enables us to explain whether or not two attributes are associated. For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not, χ^2 test will helps us in deciding this issue. In such a situation, we proceed with the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever. On this basis we first calculate the expected frequencies and then work out the value of χ^2. If the calculated value of χ^2 is less than the table value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands which means that the two attributes are independent or not associated (i.e., the new medicine is not effective in controlling the fever). But if the calculated value of χ^2 is greater than its table value, our inference then would be that null hypothesis does not hold good which means the two attributes are associated and the association is not because of some chance factor but it exists in reality (i.e., the new medicine is effective in controlling the fever and as such may be prescribed). It may, however, be stated here that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes, but is simply a technique of judging the significance of such association or relationship between two attributes.

In order that we may apply the chi-square test either as a test of goodness of fit or as a test to judge the significance of association between attributes, it is necessary that the observed as well as theoretical or expected frequencies must be grouped in the same way and the theoretical distribution must be adjusted to give the same total frequency as we find in case of observed distribution. χ^2 is then calculated as follows:

$$\chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

O_{ij} = observed frequency of the cell in *i*th row and *j*th column.

E_{ij} = expected frequency of the cell in *i*th row and *j*th column.

If two distributions (observed and theoretical) are exactly alike, $\chi^2 = 0$; but generally due to sampling errors, χ^2 is not equal to zero and as such we must know the sampling distribution of χ^2 so that we may find the probability of an observed χ^2 being given by a random sample from the hypothetical universe. Instead of working out the probabilities, we can use ready table which gives probabilities for given values of χ^2. Whether or not a calculated value of χ^2 is significant can be

ascertained by looking at the tabulated values of χ^2 for given degrees of freedom at a certain level of significance. If the calculated value of χ^2 is equal to or exceeds the table value, the difference between the observed and expected frequencies is taken as significant, but if the table value is more than the calculated value of χ^2, then the difference is considered as insignificant i.e., considered to have arisen as a result of chance and as such can be ignored.

As already stated, degrees of freedom* play an important part in using the chi-square distribution and the test based on it, one must correctly determine the degrees of freedom. If there are 10 frequency classes and there is one independent constraint, then there are (10 – 1) = 9 degrees of freedom. Thus, if '*n*' is the number of groups and one constraint is placed by making the totals of observed and expected frequencies equal, the d.f. would be equal to (*n* – 1). In the case of a contingency table (i.e., a table with 2 columns and 2 rows or a table with two columns and more than two rows or a table with two rows but more than two columns or a table with more than two rows and more than two columns), the d.f. is worked out as follows:

$$\text{d.f.} = (c - 1)(r - 1)$$

where '*c*' means the number of columns and '*r*' means the number of rows.

Example 1:

A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Solution: Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining any one of the six numbers is 1/6 and as such the expected frequency of any one number coming upward is 132 ×1/6 = 22. Now we can write the observed frequencies along with expected frequencies and work out the value of χ^2 as follows:

No. turned up	*Observed frequency* O_i	*Expected frequency* E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	16	22	–6	36	36/22
2	20	22	–2	4	4/22
3	25	22	3	9	9/22
4	14	22	–8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

$\therefore \quad \Sigma [(O_i - E_i)^2/E_i] = 9.$

Hence, the calculated value of $\chi^2 = 9$.

$\because$ Degrees of freedom in the given problem is

$(n - 1) = (6 - 1) = 5$.

The table value* of χ^2 for 5 degrees of freedom at 5 per cent level of significance is 11.071. Comparing calculated and table values of χ^2, we find that calculated value is less than the table value and as such could have arisen due to fluctuations of sampling. The result, thus, supports the hypothesis and it can be concluded that the die is unbiased.

Example 2:

Find the value of χ^2 for the following information:

Class	*A*	*B*	*C*	*D*	*E*
Observed frequency	8	29	44	15	4
Theoretical (or expected) frequency	7	24	38	24	7

Solution: Since some of the frequencies less than 10, we shall first re-group the given data as follows and then will work out the value of χ^2 :

Class	*Observed frequency* O_i	*Expected frequency* E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
A and *B*	(8 + 29) = 37	(7 + 24) = 31	6	36/31
C	44	38	6	36/38
D and *E*	(15 + 4) = 19	(24 + 7) = 31	–12	144/31

$\therefore \quad \chi^2 = \Sigma \dfrac{(O_i - E_i)^2}{E_i} = 6.76$ app.

Example 3:

Genetic theory states that children having one parent of blood type *A* and the other of blood type *B* will always be of one of three types, *A*, *AB*, *B* and that the proportion of three types will on an average be as 1 : 2 : 1. A report states that out of 300 children having one *A* parent and *B* parent, 30 per cent were found to be types *A*, 45 per cent per cent type *AB* and remainder type *B*. Test the hypothesis by χ^2 test.

Solution: The observed frequencies of type *A*, *AB* and *B* is given in the question are 90, 135 and 75 respectively.

The expected frequencies of type *A*, *AB* and *B* (as per the genetic theory) should have been 75, 150 and 75 respectively.

We now calculate the value of χ^2 as follows:

Type	*Observed frequency* O_i	*Expected frequency* E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
A	90	75	15	225	225/75 = 3
AB	135	150	–15	225	225/150 = 1.5
B	75	75	0	0	0/75 = 0

$$\therefore \qquad \chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = 3 + 1.5 + 0 = 4.5$$

$$\because \qquad \text{d.f.} = (n - 1) = (3 - 1) = 2.$$

Table value of χ^2 for 2 d.f. at 5 per cent level of significance is 5.991.

The calculated value of χ^2 is 4.5 which is less than the table value and hence can be ascribed to have taken place because of chance. This supports the theoretical hypothesis of the genetic theory that on an average type *A*, *AB* and *B* stand in the proportion of 1 : 2 : 1.

Example 4:

The following information is obtained concerning an investigation of 50 ordinary shops of small size:

	Shops		*Total*
	In towns	*In villages*	
Run by men	17	18	35
Run by women	3	12	15
Total	20	30	50

Can it be inferred that shops run by women are relatively more in villages than in towns? Use χ^2 test.

Solution: Take the hypothesis that there is no difference so far as shops run by men and women in towns and villages. With this hypothesis the expectation of shops run by men in towns would be:

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N}$$

where A = shops run by men
B = shops in towns
$(A) = 35$; $(B) = 20$ and $N = 50$

$$\text{Thus, expectation of } (AB) = \frac{35 \times 20}{50} = 14$$

Hence, table of expected frequencies would be

	Shops in towns	*Shops in villages*	*Total*
Run by men	14 (*AB*)	21 (*Ab*)	35
Run by women	6 (*aB*)	9 (*ab*)	15
Total	20	30	50

Calculation of χ^2 value:

Groups	*Observed frequency* O_{ij}	*Expected frequency* E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2/E_{ij}$
(*AB*)	17	14	3	9/14 = 0.64
(*Ab*)	18	21	–3	9/21 = 0.43
(*aB*)	3	6	–3	9/6 = 1.50
(*ab*)	12	9	3	9/9 = 1.00

$$\therefore \qquad \chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 3.57$$

As one cell frequency is only 3 in the given 2 × 2 table, we also work out χ^2 value applying Yates' correction and this is as under:

$$\chi^2 \text{(corrected)} = \frac{[|17 - 14| - 0.5]^2}{14} + \frac{[|18 - 21| - 0.5]^2}{21} + \frac{[|3 - 6| - 0.5]^2}{6} + \frac{[|12 - 9| - 0.5]^2}{9}$$

$$= \frac{(2.5)^2}{14} + \frac{(2.5)^2}{21} + \frac{(2.5)^2}{6} + \frac{(2.5)^2}{9}$$

$$= 0.446 + 0.298 + 1.040 + 0.694$$

$$= 2.478$$

$\because$ Degrees of freedom = $(c - 1)\,(r - 1) = (2 - 1)\,(2 - 1) = 1$

Table value of χ^2 for one degree of freedom at 5 per cent level of significance is 3.841. The calculated value of χ^2 by both methods (i.e., before correction and after Yates' correction) is less than its table value. Hence the hypothesis stands. We can conclude that there is no difference between shops run by men and women in villages and towns.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – III – OPERATIONS RESEARCH – SMTA5205

1. INTRODUCTION

1.1 RESEARCH ORIGIN OF OPERATIONS RESEARCH (OR)

The term Operations Research (OR) was first coined by MC Closky and Trefthen in 1940 in a small town, Bowdsey of UK. The main origin of OR was during the second world war – The military commands of UK and USA engaged several inter-disciplinary teams of scientists to undertake scientific research into strategic and tactical military operations. Their mission was to formulate specific proposals and to arrive at the decision on optimal utilization of scarce military resources and also to implement the decisions effectively. In simple words, it was to uncover the methods that can yield greatest results with little efforts. Thus it had gained popularity and was called " An art of winning the war without actually fighting it."

The name Operations Research (OR) was invented because the team was dealing with research on military operations. The encouraging results obtained by British OR teams motivated US military management to start with similar activities. The work of OR team was given various names in US: Operational Analysis, Operations Evaluation, Operations Research, System Analysis, System Research, Systems Evaluation and so on. The first method in this direction was simplex method of linear programming developed in 1947 by G.B Dantzig, USA. Since then,

new techniques and applications have been developed to yield high profit from least costs. Now OR activities has become universally applicable to any area such as transportation, hospital management, agriculture, libraries, city planning, financial institutions, construction management and so forth. In India many of the industries like Delhi cloth mills, Indian Airlines, Indian Railway, etc are making use of OR techniques.

1.2 HISTORY OF OR

The term OR coined by Mc.Clostcy and Tref in the year 1940 in U.K. OR was first used in military operations for optimum utilization of resources.

YEAR	EVENTS
1940	Term OR was coined by Mc.Closky and Trefthen in U.K
1949	OR unit was set up in India in Hyderabad. (The Regional Research Lab) OR unit was set up at defence science lab.
1951	The National Research Council (NRC) in US formed a committee on OR. The first book was published called “Methods on OR” by Morse and Kimball.
1952	OR Society of America was formed.
1953	OR unit was set up in Calcutta in the “Indian Statistical Institute”.
1995	OR society of India was established.

OR gained its significance first in the defence during the World War II (1939-1945) in order to make the best use of limited military resources and win the war. The effectiveness of OR in defence spread interest in Government departments and industry.

1.3 CONCEPT AND DEFINITION OF OR

Operations research signifies research on operations. It is the organized application of modern science, mathematics and computer techniques to complex military, government, business or industrial problems arising in the direction and management of large systems of men, material, money and machines.

The purpose is to provide the management with explicit quantitative understanding and assessment of complex situations to have sound basics for arriving at best decisions. Operations research seeks the optimum state in all conditions and thus provides optimum solution to organizational problems.

DEFINITION

"OR is defined as the application of Scientific methods, tools and techniques to problems involving the operations of a system so as to provide to those in control of the system, with optimum solutions to the problem".

"OR is defined as the application of Scientific methods by interdisciplinary team to problems involving control of organized system, so as to provide solutions which serve best to the organization as a whole."

OR is, otherwise, called as the "Science of use".

OR is the combination of management principles and mathematical concepts (Quantitative techniques) for managerial decision-making purpose.

1.4 CHARACTERISTICS OF OR

- Aims to find solutions for problems of organized systems.
- Aims to provide optimum solution. Optimization means the best minimum or maximum for the criteria under consideration.
- It is the application of scientific methods, tools and techniques.
- Interdisciplinary team approach is used to solve the problems.
- The solutions that serve best to the organization as a whole is taken into consideration.

1.5 APPLICATION OF OR

1. **Production:**
   - Production scheduling
   - Project scheduling
   - Allocation of resources
   - Equipment replacement
   - Inventory policy
   - Factory size and location
2. **Marketing**
   - Product introduction with timing
   - Product mix selection
   - Competitive strategies
   - Advertising strategies
   - Pricing strategies.
3. **Accounts**
   - Cash flow analysis (optimum cash balance)
   - Credit policies (optimum receivables)
4. **Finance**
   - Optimum dividend policy
   - Portfolio analysis
5. **Personnel Management**
   - Recruitment and selection
   - Assignment of jobs
   - Scheduling of training programs

6. **Purchasing**
   - Rules for purchasing
   - EOQ-Economic Order Quantity (how much to order)
   - Timing of purchase (when to purchase)
7. **Distribution**
   - Deciding number of warehouses.
   - Location of warehouses
   - Size of warehouses
   - Transportation strategies
8. **Defence**
   - Budget allocation
   - Allocation of resources
9. **Government Departments**
   - Transportation
   - Budget fixation
   - Fiscal policies.
10. **R & D (Research and Development)**
    - Project introduction
    - Project control
    - Budget allocation for projects

1.6 THE MAIN PHASES OF OR

- Formulation of the problem
- Construction of a model (Mathematical model)
- Solve the model
- Control and update the model
- Test the model and validate it
- Implement the model

1.6.1. FORMULATION OF THE PROBLEM

- The first task is to study the relevant system and develop a well-defined statement of the problem. This includes determining appropriate objectives, constraints, interrelationships and alternative course of action.
- The OR team normally works in an advisory capacity . The team performs a detailed technical analysis of the problem and then presents recommendations to the management.
- Ascertaining the appropriate objectives is very important aspect of problem definition. Some of the objectives include maintaining stable price, profits, increasing the share in market, improving work morale etc.
- OR team typically spends huge amount of time in gathering relevant data.
  - To gain accurate understanding of problem
  - To provide input for next phase.
- OR teams uses Data mining methods to search large databases for interesting patternsthat may lead to useful decisions.

1.6.2. CONSTRUCTION OF A MATHEMATICAL MODEL

This phase is to reformulate the problem in terms of mathematical symbols and expressions. The mathematical model of a business problem is described as the system of equations and related mathematical expressions. Thus

1. Decision variables(x1, x2… xn) – ‘n’ related quantifiable decisions to be made.
2. Objective function– measure of performance (profit) expressed as mathematical function of decision variables. For example P=3x1+5x2+ … + 4xn
3. Constraints– any restriction on values that can be assigned to decision variables in terms of inequalities or equations. For example x1+2x2≥ 20
4. Parameters– the constant in the constraints (right hand side values)The advantages of using mathematical models are

- Describe the problem more concisely
- Makes overall structure of problem comprehensible
- Helps to reveal important cause-and-effect relationships
- Indicates clearly what additional data are relevant for analysis
- Forms a bridge to use mathematical technique in computers to analyze

1.6.3. SOLVE THE MODEL

This phase is to develop a procedure for deriving solutions to the problem. A common theme isto search for an optimal or best solution. The main goal of OR team is to obtain an optimalsolution which minimizes the cost and time and maximizes the profit

1.6.4. TESTING THE MODEL

After deriving the solution, it is tested as a whole for errors if any. The process of testing and improving a model to increase its validity is commonly referred as Model validation. The OR group doing this review should preferably include at least one individual who did not participate in the formulation of model to reveal mistakes. A systematic approach to test the model is to use Retrospective test. This test uses historical data to reconstruct the past and then determine the model and the resulting solution. Comparing the effectiveness of this hypothetical performance with what actually happened, indicates whether the model tends to yield a significant improvement over current practice.

1.6.5. IMPLEMENTATION

The last phase of an OR study is to implement the system as prescribed by the management. The success of this phase depends on the support of both top management and operating management. The implementation phase involves several steps

1. OR team provides a detailed explanation to the operating management
2. If the solution is satisfied, then operating management will provide the explanation to the personnel, the new course of action.
3. The OR team monitors the functioning of the new system
4. Feedback is obtained
5. Documentation

1.7 SCOPE OF OR

- Linear programming model
- Transportation
- Sequencing and scheduling
- Assignment of jobs to minimize cost or maximize profit
- Game theory
- Inventory model
- Maintenance and Replacement
- Waiting line models
- Network analysis
- Shortest route problems like traveling sales person problem
- Resource allocation problems

1.8 LIMITATIONS OF OR

1. **Magnitude of computation:** In order to arrive at an optimum solutions OR takes into account all the variables that affect the system. Hence the magnitude of computation is very large.

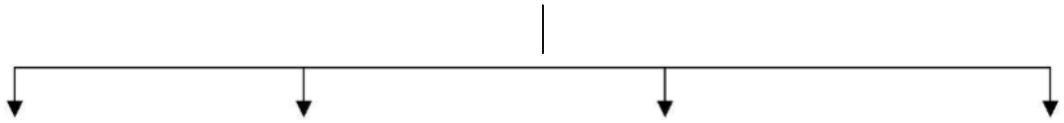
2. **Non-Quantifiable variables:** OR can give an optimum solution to a problem only if all the variables are quantified. Practically all variables in a system cannot be quantified.
3. **Time and Cost:** To implement OR in an organization, it consumes more time and cost. If the basic decision variables change, OR becomes too costly for an organization to handle it.
4. **Implementation of OR:** Implementation of OR may lead to HR problems. The psychology of employees should be considered and the success of OR depends on co-operation of the employees.
5. **Distance between Manager and OR Specialist:** Managers may not be having a complete overview of OR techniques and has to depend upon an OR Specialist. Only if good link is established OR can be a success.

2. MODELS IN OR

Model is a reasonably simplified representation of reality. It is an abstraction of reality. It helps to arrive at a well-structured view of reality.

2.1 TYPES OF MODELS

I -BASED ON NATURE



Iconic Model	Analogue Model (or) Schematic model	Mathematical Model (or) Symbolic model	Quantitative Model

- **ICONIC MODELS:**
  - It is a pictorial representation or a physical representation of a system. A look alike correspondence is present.

Eg: miniature of a building, toys, globe etc.

- Iconic Models are either scaled up or scaled down. Scaled up - eg: Atom. Scaled down – eg: globe.
- Iconic models are either two-dimensional or three-dimensional.

- **ANALOGUE MODEL OR SCHEMATIC MODEL**

This model uses one set of properties to describe another set of properties. There is no look alike correspondence. It is more abstract.

Eg: a set of water pipes that are used to describe current flow. Eg: Maps, (different colors are used to depict water, land etc Eg: Organizational chart.

- **MATHEMATICAL MODEL**

This uses a set of mathematical symbols (letters and numbers) to represent a system.

V = I * R ⟶ (Resistance)

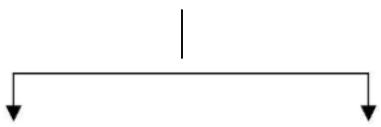
↓ ↓

(Voltage)(Current)

- **QUANTITATIVE MODELS**

Quantitative models are those, which can measure the observation. Eg: Models that measure temperature.

II -BASED ON VARIABILITY



STATIC **DYNAMIC**

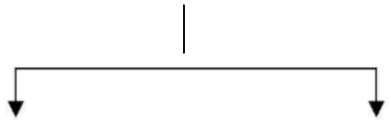
- **STATIC MODEL**

  This model assumes the values of the variables to be constant (do not change with time) eg: Assignment Model.

- **DYNAMIC MODEL**

  This model assumes that the values of the variable change with time. Eg: Replacement model.

III -BASED ON RISK COSIDERATION



DETERMINISTIC **STOCHASTIC (PROBABILISTIC)**

- **DETERMINISTIC MODEL**

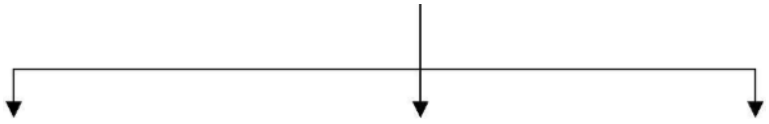
  This model does not take uncertainty into consideration.

  Eg: Linear programming, Assignment etc

- **STOCHASTIC (PROBABILISTIC) MODEL**

  This model considers uncertainty as an important factor. Eg: Stochastic Inventory models.

IV -BASED ON PRESENTATION



DESCRIPTIVE **PREDICTIVE** **PRESCRIPTIVE**

- **DESCRIPTIVE MODEL**

  This model just describes the situation under consideration. Eg: collecting an opinion regarding tendency to vote.

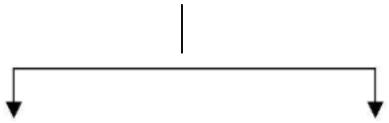
- **PREDICTIVE MODEL**

  This is a model, which predicts the future based on the data collected. Eg: predicting the election results before actual counting.

- **PRESCRIPTIVE MODEL**

  This is a model, which prescribes the course of action to be followed. Eg: Linear programming.

V -BASED ON ACCURACY OR EXACTNESS



ANALYTICAL **SIMULATION**

- **ANALYTICAL MODEL**

  This is a model that gives an exact solution to the problem.

- **SIMULATION MODEL**

  Simulation model is a representation of reality through the use of some devices, which will react in the same manner under the given set of conditions.

  Eg: Simulation of a drive of an Airplane through computer.

3. LINEAR PROGRAMMING

3.1 DEFINITION

Samuelson, Dorfman, and Solow define LP as "the analysis of problems in which linear function of a number of variables is to be maximized (or minimized) when those variables are
subject to a number of constraints in the form of linear inequalities".

3.2 BASIC ASSUMPTIONS OF LINEAR PROGRAMMING:

The following four basic assumptions are necessary for all linear programming models:

1. LINEARITY:

The basic requirements of a LP problem are that both the objectives and constraints must be expressed in terms of linear equations or inequalities. It is well known that if the number of machines in a plant is increased, the production in the plant also proportionately increases. Such a relationship, giving corresponding increment in one variable for every increment in other, is called linear and can be graphically represented in the form of a straight line.

2. DETERMINISTIC (OR CERTAINTY):

In all LP models, it is assumed that all model parameters such as availability of resources, profit (or cost) contribution of a unit of decision variable and consumption of resources by a unit decisions variable must be known and fixed. In other words, this assumptions means that all the **coefficients** in the objectives function and constraints are completely known with certainty and do not change during the period being studied.

3. ADDITIVITY:

The value of the objective function for the given values of decision variables and the total sum of resources used, must be equal to the sum of the contributions (profit or cost) earned from each decision variable and the sum of the resources used by each decision respectively. For example, the total profit earned by the sale of three products A, B and C must be equal to the profits earned separately from A, B and C and similarly, the amount of

resources consumed by A, B, and C individually.

4. DIVISIBILITY:

This implies that solution values of decision variables and resources can take any non-negative values, i.e., **fractional values** of the decision variables are **permitted**. This, however, is not always desirable. For example, it is impossible to produce one-fourth of a bus. When it is necessary to have integer variables, a technique known as integer programming could be used.

3.4 APPLICATIONS OF LINEAR PROGRAMMING:

(i) **MANUFACTURING PROBLEMS:** to find the number of items of each type that should be manufactured so as to maximize the profit subject to production restrictions imposed by limitations on the use of machinery and labour.

(ii) **ASSEMBLING PROBLEMS:** To have the best combinations of basic components to produce goods according to certain specifications.

(iii) **TRANSPORTATION PROBLEMS**: to find the least costly way of transporting shipments from the warehouses to customers.

(iv) **BLENDING PROBLEM:** To determine the optimal amount of several constitutes to use in producing a set of products which determining the optimal quantity of each product to produce.

(v) **PRODUCTION PROBLEMS:** To decide the production schedule to satisfy demand and minimize cost in face of fluctuating rates and storage expenses.

(vii) **DIET PROBLEMS**: To determine the minimum requirement of nutrients subject to availability of foods and their prices.

(viii) **JOB ASSIGNING PROBLEMS:** To assign job to workers for maximum effectiveness and optimal results subject to restrictions of wages and other costs.

(ix) **TRIM-LOSS PROBLEMS:** To determine the best way to obtain a variety of smaller rolls of paper from a standard width of roll that it kept its stock and at the same time minimize wastage.

(x) **STAFFING PROBLEM**: To find optimal staff in hotels, police stations and hospitals to maximize the efficiency.

(xi) **TELEPHONE EXCHANGE PROBLEMS:** To determine optimal facilities in telephone exchange to have minimum breakdowns.

3.5 APPLICATIONS OF LINEAR PROGRAMMING

a) Personnel Assignment Problem
b) Transportation Problem
c) Efficiency on Operation of system of Dams
d) Optimum Estimation of Executive Compensation
e) Agriculture Applications
f) Military Applications
g) Production Management
h) Marketing Management
i) Manpower Management

3.6 KEY TERMS

Artificial variables: A variable that has no meaning in a physical sense, but acts as tool to help generate an initial LP solution.

Basic variables: The set of variables that are in the solution (i.e., have positive, on-zero values) are listed in the product mix column. The variables that normally take non-zero values to obtain a solution.

Basic solution: A solution to m simultaneous linear equations in n unknowns, m<n, with the property that n-m of the variables have the value zero and the values of the remaining m variables are unique determined; obtained when a set of non-basic variables are assigned the

value zero.

Basic feasible solution: A basic solution, for which the values of all variables are non-negative, corresponds to a corner of the LP feasible region.

Degeneracy : A condition that arises when there is a tie in the values used to determine which variables indicated will enter the solution next. It can lead to cycling back and forth between two non-optimal solutions.

Degenerate solution: The number of variables in the standard equality form (counting decision variables, surpluses, and slacks) with positive optimal value is less than the number of constraints.

Optimal solution: A solution that is optimal for the given solution.

Pivot column: The column with the largest positive number in the Ci-Zj row of a maximization problem, or the largest negative Cj-Zj value in a minimization problem. It indicates which variable will enter the solution next.

Pivot row: The corresponding to the variable that will leave the basis in order to make room for the variable entering (as indicated by the new pivot column). This is the smallest positive ratio found by dividing the quantity column values by the pivot column values for each row.

Slack variable: A variable added to less than or equal to constraints in order to create an equality for a simplex method. It represents a quantity of unused resources.

Surplus variable: A variable inserted in a greater than or equal to constraint to create equality. If represents the amount of resources usage above the minimum required usage.

Unboundedness: A condition describing LP maximization problems having solutions that can become infinitely large without violating any stated constraints.

3.7 ADVATNTAGES OF LPP:

- It provides an insight and perspective into the problem environment. This generally results in clear picture of the true problem.
- It makes a scientific and mathematical analysis of the problem situations.
- It gives an opportunity to the decision-maker to formulate his strategies consistent with the constraints and the objectives.
- It deals with changing situations. Once a plan is arrived through the LP it can also be

revaluated for changing conditions.

- By using LP, the decision maker makes sure that he is considering the best solution.

3.8 LIMITATIONS OF LPP:

- The major limitation of LP is that it treats all relationships as linear but it is not true in many real life situations.
- The decision variables in some LPP would be meaningful only if they have integer values. But sometimes we get fractional values to the optimal solution, where only integer values are meaningful.
- All the parameters in the LP model are assumed to be known constants. But in real life they may not be known completely or they may be probabilistic and they may be liable for changes from time to time.
- The problems are complex if the number of variables and constraints are quite large.
- It deals with only single objective problems, whereas in real life situations, there may be more than one objective.

3.9 FORMULATION OF LPP:

- Identify the objective function
- Identify the decision variables
- Express the objective function in terms of decision variables
- Identify the constraints and express them
- Value of decision variables is ≥ 0 (always non-negativity)

EXAMPLE PROBLEM:

An organization wants to produce Tables and Chairs. Profit of one table is ₹100 and profit of one Chair is ₹50

Particulars	Tables	Chairs	Maximum hours available
Cutting (hours)	4	1	300
Painting (hours)	1	.5	100

. . . . .

. . . . .

. . . . .

Solution:

1) Objective: Maximization of profit
2) Decision variables

   No. of Tables to be produced ‘x’

   No. of Chairs to be produced ‘y’
3) Objective function Maxi Z = 100x + 50y
4) Constraints 4x +

   1y ≤ 300

   1x + 0.5y ≤ 100 x ≥

   0, y ≥ 0
5) Formulate

   Maxi Z = 100x + 50y

   Subject to

   4x + 1y ≤ 300

   1x + 0.5y ≤ 100 x, y ≥ 0

3.9 STEPS IN GRAPHICAL SOLUTION METHOD:

- Formulate the objective and constraint functions.
- Draw a graph with one variable on the horizontal axis and one on the vertical axis.
- Plot each of the constraints as if they are inequalities.
- Outline the solution area.
- Circle the potential solutions points. These are the intersections of the constraints on the perimeter of the solution area. (vertices of the solution space)
- Substitute each of the potential extreme point values of the two decision variables into the objective function and solve for Z.
- Select the solution that optimizes Z.

3.10 PROCEEDURE FOR SOLVING LPP PROBLEM USING SIMPLEX METHOD

STEP:1

Convert all the inequality functions into equality:

For converting all the inequalities into equalities, we should use slack and surplus variables.

In case of ≤ inequalities, we should add Slack variable so as to convert that inequality into equality. For example, $3x + 2y \leq 6$ will become $3x + 2y + S1 = 6$, where S1 is the slack variable.

In case of ≥ inequalities, we should deduct Surplus variable

so as to convert that inequality into equation.

For example, $5x + 6y \geq 10$ will become $5x + 6y – S2 = 10$, where S2 is the surplus variable.

In case if the given constraint is an equation category, we should not use either slack variable or surplus variable.

STEP 2:

Find out the basic and non basic variables: Non Basic variable is the variable whose value is zero. Basic variable is the variable which will have either positive or negative value.

After converting all the inequality into equality, we should assume some variables as Non basic variables and find out the values of the other (Basic) variables. This solution is called as initial solution. If all the basic variable values are positive, then that initial solution is called as BASIC FEASIBLE SOLUTION.

STEP:3

Preparation of simplex table: The format of the simplex table is as follows:
Coefficients of Variables in the Objective function

EVALUATION ROW

Coefficients of Basic Variables	Basic Variables	Variables	Solution	Ratio

STEP 4:

Calculation of values in Evaluation row: To calculate the values in the evaluation row, we should use the following formula for each variable column:

Evaluation row values = (Variable coefficients x coefficients of basic variables) – Coefficients of the variables in the objective function.

All the values in the Evaluation row should be either positive or zero. Then it indicates that we have reached the optimum stage and thereby we can derive the optimum solution.

If any negative persists, we should proceed further by doing the following steps.

STEP 5:

IDENTIFICATION OF KEY COLUMN: The column that represents least value in the evaluation row is known as KEY COLYMN. The variable in that column is known as ENTERING VARIABLE.

STEP 6:

IDENTIFICATION OF KEY ROW: To find out the Key row, we should calculate the ratio.

Ratio = solution column values / Key column values. The least ratio row is treated as KEY ROW and the value in that row is known as LEAVING VARIABLE. **THE VARIABLE THAT**

PREVAILS IN BOTH KEY ROW AND KEY COLUMN IS KNOWN AS KEY ELEMENT. After finding the key element, we should prepare next simplex table. In that table, should bring the entering variable and should write the new values of the entering

Problems

Problem 1. A person requires 10, 12 and 12 units of chemicals A, B and C respectively for herbal garden. A liquid product contains 5, 2 and 1 units of A,B and C respectively per Jar. A dry product contains 1, 2 and 4 units of A, B and C per cartoon. If the liquid product sells for Rs. 3 per Jar and dry product sells for Rs. 2 per cartoon, how many of each should be purchased to minimise the cost and meet the requirements.

Solution :

Requirement	A	B	C	
	10	12	12 units	
Liquid product	5	2	1	units Rs. 3/-per jar
Dry product	1	2	4	Rs. 2/-per Cartons

1. Select decision variable

   x_1 – no. of jars of liquid product

   x_2 – no. of cartoons of dry product

2. Objective function

   Minimize cost $(z) = 3x_1 + 2x_2$

3. Constraints :

$$5x_1 + x_2 \geq 10$$
$$2x_1 + 2x_2 \geq 12$$
$$1x_1 + 4x_2 \geq 12$$

4. Add non negativity constraints :

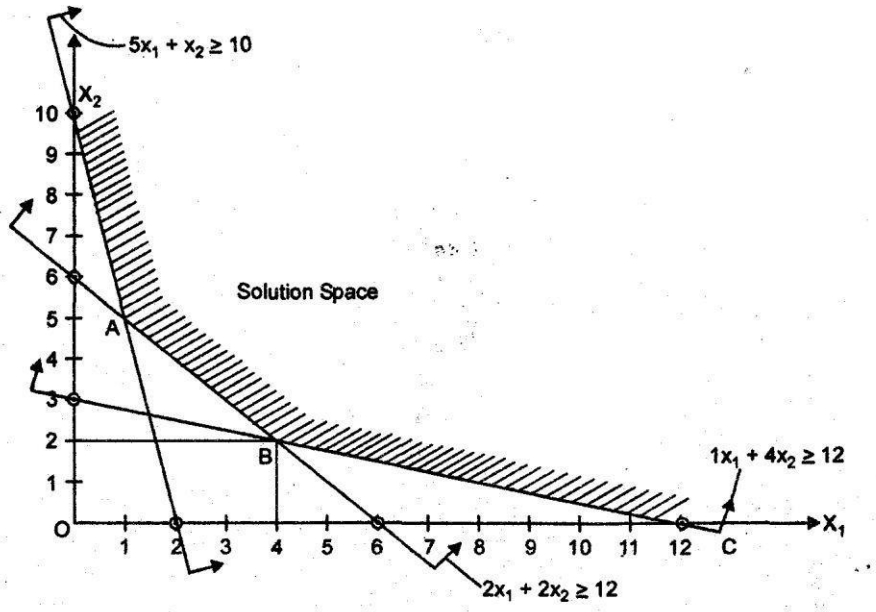
$$x_1 \geq 0 \quad ; \quad x_2 \geq 0$$

Graphical Method :

$5x_1 + x_2 = 10 \Rightarrow x_1 = 0;\ x_2 = 10$ and $x_2 = 0;\ x_1 = 2$

$2x_1 + 2x_2 = 12 \Rightarrow x_1 = 0;\ x_2 = 6$ and $x_2 = 0;\ x_1 = 6$

$1x_1 + 4x_2 = 12 \Rightarrow x_1 = 0;\ x_2 = 3$ and $x_2 = 0;\ x_1 = 12$



Point A (1,5) $Z(A) = 3 \times 1 + 2 \times 5 = 13$

Point B (4, 2) $Z(B) = 3 \times 4 + 2 \times 2 = 16$

Point C (12,0) $Z(C) = 3 \times 12 + 2 \times 0 = 36$

Minimum cost at point A *i.e.* Rs. 13

x_1 (no. of Jar of Liquid product) = 1

x_2 (no. of carton of dry product) = 5

Minimum cost (Z) = Rs. 13.

Problem 2. A firm manufactures pain relieving pills in two sizes A and B, size A contains 4 grains of element a, 7 grains of element b and 2 grains of element c, size B contains 2 grains of element a, 10 grains of element b and 8 grains of c. It is found by users that it requires at least 12 grains of element a, 74 grains of element b and 24 grains of element c to provide immediate relief It is required to determine that least no. of pills a patient should take to get immediate relief. Formulate the problem as standard LPP.

Solution : Pain relieving pills

	a „	.b	e
Size A	4	7	2
Size B	2	10	8
Min. requirement	12	74	24

Step 1. Select decision variable

z_1 - no. of pills of size A

x_2 - no. of pills of size B

Step 2. Objective function

Minimum (no. of pills) z = z$+ z_2

Step 3. Constraints

7 > 10z_2 74

2zt + 8z_2y 24

Step. 4. Add non negativity constraints

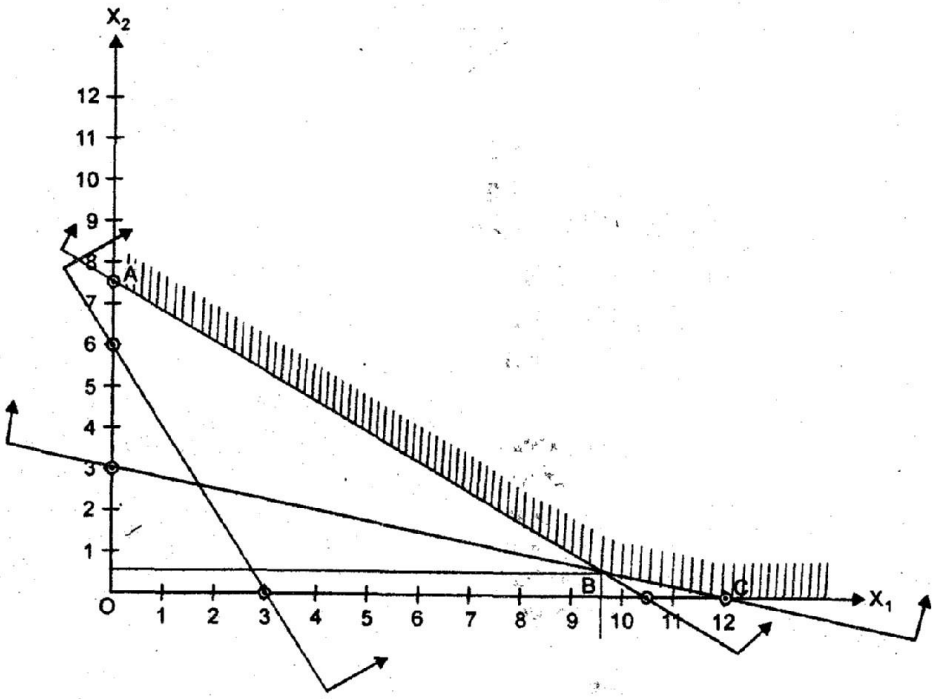
z_1 ¿ 0 ; x 0

Determining the value of z and x by graphical method

4x$+ 2xq = 12 z$ =0; xy= 6 and xy = 0; xy=3

7z$* J0$z_2$ = 74 z = 0; z_2= 7’4 and zg- 0r; $_d$ = 10.57

2zy+ 8x_2 = 24 z, = a; x_2 = 5 and x = 0; zp = t2



Point A (0, 7.4) Z(A) = 0 + 7.4 = 7.4 (Minimum)
Point C (12, 0) Z (C) = 12 + 0 = 12
Point B (9.6, 0.6) Z(B) = 9.6 + 0.6 = 10.2
No. of pills of size A = 0
No. of pills of size B = 7.4 ≃ 8 pills
Minimum no. of pills = 8 pills.

Problem 3. An automobile manufactuEer makes automobiles and trucks in a factory that is divided into two shops. Shop A which perform the basic assy operation must work 5 man days on each truck but only 2 man days on each automobile. Shop B which perform finishing operations must work 3 man days for each automobile or truck that it produces. Because of men and machine limitations shop A has 180 man days per week available while shop B has 135.man days per week. If the manufacturer makes a profit of Rs. 300 on each truck and Rs. 200 on each automobile; how many of each should be produced to maximize his profit?

Solution :

	Shop A	Shop B	Profit
Automobile	2 man days	3 man days	Rs. 200
Trucks	5 man days	3 man days	Rs. 300
Availability	180 man days/week	135 man days/week	

1. Select decision variable
   z - no. of automobile to be produced/week
   r$ — no. of trucks to be produced/week
2. Objective function
   Maximize Z = 200a + 300x_2
3. Constraints

$$2x_1 + 5x_2 \leq 180$$
$$3 + 3z_2 \quad 135$$

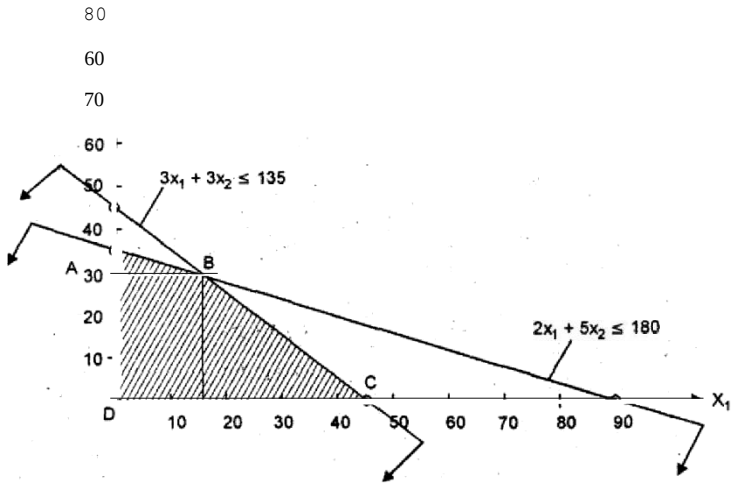
4. Add non negativity constraints

z 0; zq 0

Determine the value of z and z_2 by graphical method

$2x_1+5x_2 = 180$ x - *0*, z_2= 36 and xg = 0; x = 90

$3x_1+3x_2 = 135$ x = 0; z_2= 45 and x_2 = 0; z = 45



Point D (0, 0) Z (D) $= 200 \times 0 + 300 \times 0 = 0$

Point A (0, 36) Z (A) $= 200 \times 0 + 300 \times 36 = 10800$

Point C (45, 0) Z (C) $= 200 \times 45 + 300 \times 0 = 9000$

Point B (15 , 30) Z (B) $= 200 \times 15 + 300 \times 30 = 3000 + 9000 = 12000$

Maximum Profit at Point B (15 , 30) *i.e.* Rs.12000/-

x_1 = no. of automobile/week = 15

x_2 = no. of trucks/week = 30

Maximum profit = 12000/-

Problem 4. On completing the construction of house a person discovers that J square feet of plywood scrap and 80 square feet of white pine scrap are in use!able m for the construction of tables and book cases. It takes 16 square feet of plywood 8 square feet of white pine to make a table, 12 square feet of plywood and 16 Llare feet of white pine are required to contruct a book case. By selling the finishing duct to a local furniture store the person can realize a profit of Rs. 25 on each table d Rs. 290 on each book case. How may the man most profitably use the left over ood ? Use graphical method to solve problem.

Solution :

	Plywood	White pine	Profit	
Table	16	8	Rs. 25	each table
Book case	12	16	Rs. 290	each book case
Availability	100	80		

1. Select decision variable

   x_1 – no. of table

   x_2 – no. of book case

2. Objective function

   Maximize profit (Z) $= 25x_1 + 290x_2$

3. Constraints

$$16x_1 + 12x_2 \leq 100$$

$$8x_1 + 16x_2 \leq 80$$

4. Add non negativity constraints

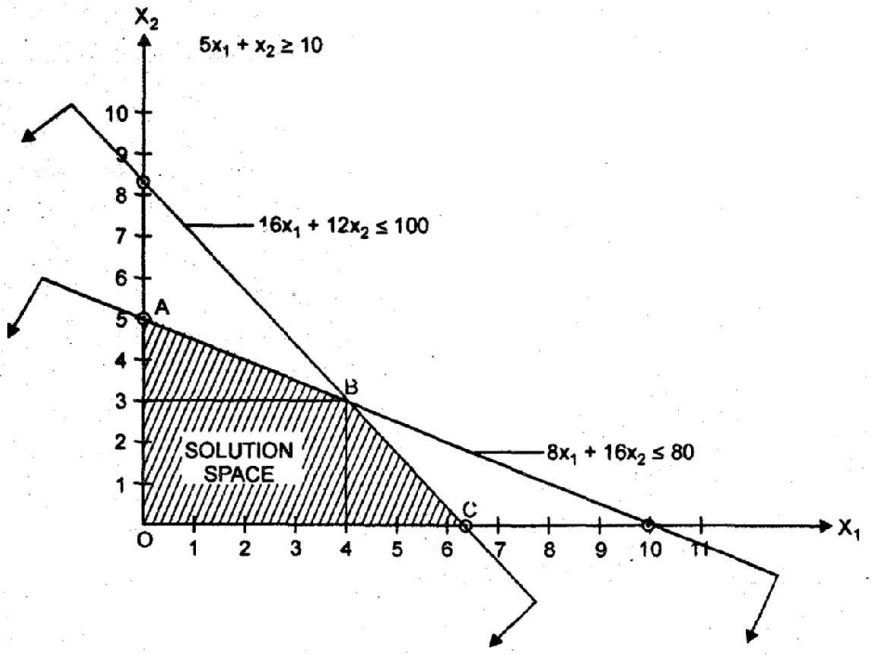
$$x_1 \geq 0$$

$$x_2 \geq 0$$

Determine the value of x_1 and x_2 using graphical method

$16x_1 + 12x_2 = 100$ $\quad x_1 = 0;\ x_2 = 8.3$ and $x_2 = 0;\ x_1 = 6.25$

$8x_1 + 16x_2 = 80$ $\quad x_1 = 0;\ x_2 = 5$ and $x_2 = 0;\ x_1 = 10$



Point O (0, 0) Z(O) = 25 × 0 + 290 × 0 = 0

Point A (0, 5) Z (A) = 25 × 0 + 290 × 5 = 1450

Point C (6.25, 0) Z (C) = 25 × 6.25 + 290 × 0 = 156.25

Point B (4, 3) Z (B) = 25 × 4 + 290 × 3 = 100 + 870 = 970

Maximum profit (Z) at point A *i.e.* Rs. 1450.

x_1 – no. of table = 0 Max. Profit (Z) = Rs. 1450/–

x_2 – no.of bookcase = 5.

Problem 5. A truck can carry a total of 10 tons of product. Three types of produts are available for shipment. Their weight and values are tabulated. Assuming that at. least one of each type must be shipped. Determine the loading which will maximize the total value. Formulate the problem.

Type	Value (Rs)	Weight (tons)
A	20	1
B	50	2
C	60	2

Solution : 1. Select decision variable

x_1 – no. of type A products

x_2 – no. of type B products

x_3 – no. of type C products

2. Objective function

Maximize (total value) $Z = 20x_1 + 50z_2 + 60x_3$

3. Constraints

$$x_1 + 2x_2 + 2x_3 \leq 10$$
$$x_1 \geq 1 \Rightarrow x_1 = 1+x$$
$$x_2 \geq 1 \Rightarrow x_2 = 1+y$$
$$x_3 \geq 1 \Rightarrow x_3 = 1+z$$

Put these values in constraint t.

(I + z) + 2 (1 + y) + 2 (t+ z) fi 10

1 + z + 2 + 2y + 2 + 2z fi IO

x + 2y * 2zs 5

4. Add non negativity constraints

x * 0; y n 0; Z n 0

Objective function in terms of ;r, y, z

Maximize (V) = 20 (1 + *x)* + 50 (1+ y} + 60 (1+z)

= 20 +20x + 50 + 50y + 60 + 60z

Max. (U) = 20a + 50y +.60a +130

Subject to

'z + 2y + 2z s 5

x >• 0; y a0; z z 0;

Simplex method

2. Find initial basic feasible solution

Substituting $z_1 = z = z =$ fi in equation (i), (ii) and (iit)

Sj

S = IG

S, =IS

3. Perform optimality test

		C	3	5	4	0	0	0		
FR	Cg	Basis	zj	„2		Sj	Sq	S	b	8
Key row	0	S	2	key element 3	0	3	0	0	fl	8/3
2/3	0	S_2	0	2	5	0	1	0	10	S
2/3	0	5_3	3	2	4	0	0	1	J5	15/2
		Z	0	0	0	0	0	0		
		C_j-Z_j	3	5 key column	4	0	0	0		

Key row

4. Iterate towards optmal solution

		C	3	5	4	0	0	U			
FR	C_B	Basis	x_1	x_2	x_3	S_1	S_2	S_3	b	θ	
	5	x_2	2/3	i	U	1/3	0	0	8/3		
	0	S	—4/3	0	5	-2/3	1	0	14/3	14/15	Key row
4/5	0	S	5/3	0	4	-2/3	0	1	29/3	29/12	
		Z_J	10/3	5	0	5/3	0	0	40/3		
		Ct—Zt	-1/3	0	4	-5/3	0	0			

Key rolumn

		Ct	3	5	4	0	0	0			
FR	C_B	Basis	z	z2	x3	S	S_2	S_3	b	8	
10/41		x_2	2/3	J	*0*	1/3	0	0	8/3	4	
$\frac{-4}{41}$	4	3	W/15	0	1	—2/15	1/5	0	14/15	-7/2	
	0	S	41/15	0	0	—2/15	—4/5	1	89/IN	89/41	Key row +-—
		Z	34/15	5	4	17/ 15	4/5	0	/15		
		CJ—ZJ	11/15	0	0	—17/JS	—4/5	0			

	Ct	3	5	4	*0*	*0*	0		
Cg	Basis	z,	z_2	z	S	S_2	S_3	b	8
5	z2	0	1	0	45/123	8/41	—10/41	50/41	
4	3	0	0	1	-6/41	33/205	4/41	62/41	
3	xy	1	0	0	-2/41	-12/41	15/41	89/41	
		3	5	4	135/41	152/41	11/41	7 /4	
	Ct-ZJ	0	0	0	-135/41	-152/41	—11/41		

Since all element of CJ -ZJ row are negative or zero so optimality test is passed

$$x_1 = \frac{89}{41}$$

2" 41

62

'3 41

$$\text{Max } z = 3 \times \frac{89}{41} + 5 \times \frac{50}{41} + 4 \times \frac{62}{41}$$

Ans. $\frac{765}{41}$

Problem 6. **Show that there is an unbounded solution to the following L.P. problem.**

$$\text{Maximize } Z = 4x_1 + x_2 + 3x_3 + 5x_4$$

$$\text{Subject to } 4x_1 - 6x_2 - 5x_3 - 4x_4 \geq -20$$

$$-3x_1 - 2x_2 + 4x_3 + x_4 \leq 10$$

$$-8x_1 - 3x_2 + 3x_3 + 2x_4 \leq 20$$

$$x_1, x_2, x_3, x_4 \geq 0$$

Solution : Standard form

Multiplying the first constraint by -1

$$-4x_1 + 6x_2 + 5x_3 + 4x_4 \leq 20$$

$$-4x_1 + 6x_2 + 5x_3 + 4x_4 + S_1 = 20$$

$$-3x_1 - 2x_2 + 4x_3 + x_4 + S_2 = 10$$

$$-8x_1 - 3x_2 + 3x_3 + 2x_4 + S_3 = 20$$

$$x_1, x_2, x_3, x_4, S_1, S_2, S_3 \geq 0$$

2. ibfs

$$x_1 = x_2 = x_3 = x_4 = 0$$

$$S_1 = 20; S_2 = 10; S_3 = 20; Z = 0$$

3. Iterations for optimal solution

	C_J	4	1	3	5	0	0	0			
C_B	Basis	x_1	x_2	x_3	x_4	S_1	S_2	S_3	b	θ	
0	S_1	-4	6	5	4	1	0	0	20	5	← Key row
0	S_2	-3	-2	4	1	0	1	0	10	10	
0	S_3	-8	-3	3	2	0	0	1	20	10	
	Z_J	0	0	0	0	0	0	0	0		
	$C_J - Z_J$	4	1	3	5	0	0	0			

↑

	C_J	4	1	3	5	0	0	0		
C_B	Basis	x_1	x_2	x_3	x_4	S_1	S_2	S_3	b	θ
5	x_4	-1	3/2	5/4	1	1/4	0	0	5	-5
0	S_2	-2	-7/2	11/4	0	-1/4	1	0	5	-5/2
0	S_3	-6	-6	1/2	0	-1/2	0	1	10	-5/3
	Z_J	-5	15/2	25/4	5	5/4	0	0	25	
	C_J-Z_J	9	-13/2	-13/4	0	-5/4	0	0		

Since all replacement ratios (θ) are negative the problem has no bounded solution and further computation stop. (Unbounded Solution)

Problem 7. **6.Maximize $Z = x_1 + 2x_2 + 3x_3 - x_4$**

Subject to $x_1 + 2x_2 + 3x_3 = 15$

$2x_1 + x_2 + 5x_3 = 20$

$x_1 + 2x_2 + x_3 + x_4 = 10$

$x_1, x_2, x_3, x_4 \geq 0$

Solution : Step 1. Standard form

Maximize $Z = x_1 + 2x_2 + 3x_3 - x_4 - MA_1 - MA_2 - MA_3$

Subject to $x_1 + 2x_2 + 3x_3 + 0x_4 + A_1 + 0A_2 + 0A_3 = 15$

$2x_1 + x_2 + 5x_3 + 0x_4 + 0A_1 + A_2 + 0A_3 = 20$

$x_1 + 2x_2 + x_3 + x_4 + 0A_1 + 0A_2 + A_3 = 10$

$x_1, x_2, x_3, x_4, A_1, A_2, A_3 \geq 0$

Step 2. ibfs

$x_1 = x_2 = x_3 = x_4 = 0$

$A_1 = 15$

$A_2 = 20$

$A_3 = 10$

$Z = -45\ M$

Step 3. Check for optimality test

	C_J	1	2	3	-1	-M	-M	-M			
C_B	Basis	x_1	x_2	x_3	x_4	A_1	A_2	A_3	b	θ	
-M	A_1	1	2	3	0	1	0	0	15	5	
-M	A_2	2	1	(5)	0	0	1	0	20	4	Key row ←
-M	A_3	1	2	1	1	0	0	1	10	10	
	Z_J	-4M	-5M	-9M	-M	-M	-M	-M	-45M		
	$C_J - Z_J$	1+4M	2+5M	3+9M	-1+M	0	0	0			

	C	1	2	3	-1	-M	-M			
Cl	Basis	z	$x_{¿}$	x	z_4			b	8	
—M	Ay	- J/5	(7/5)	0	0	1	0	3	15/7	Key now
?	z	2/5	I/5	1	0	0	0	4	20	
- M	A_3	3/5	9/5	0	1	0	1	6	10/3	
		$\frac{6-2M}{5}$	$\frac{3-16M}{s}$	3	M	M	M	12 — 9M		
	J ZJ	$\frac{-J+2M}{5}$	$\frac{7+16M}{5}$	0	- 1 + M	0	0			

	C_J	1	2	3	-1	- M			
C_B	Basis	zj	x_2	z_3	x_4	Ay	b	8	
2	z$	- 1/7	1	0	0	0	15/7	=	
3	z	3/7	0	1	0	0	25/7	«	
—M	Ay	6/7	0	0	(I)	1	15/7	15/7	> Key row
	z	$\frac{7-6M}{7}$	2	3	—M	—M	$\frac{105-JSM}{7}$		
	NJ - zJ	$\frac{6M}{y}$	0	0	—1 + M	0			

	CJ	1	2	3	—1			
C_B	Basis	=	=	‘ 3	‘4	b	*0*	
2	2	—1/7	1	0	0	15/7	-15	
3	3	3/7	0	1	0	25/7	25/3	
—1	x4	6/7	0	0	1	T5/7	5/2	Key row
	ZJ	1/7	2	3	-1	90/7		
	CJ ZJ	6 7	0	0	0			

	CJ	1	2	3	-1	
CB	Basis	x_1	x_2	x_3	x_4	b
2	x_2	0	1	0	1/6	5/2
3	x_3	0	0	1	—1/2	5/2
1	x_1	1	0	0	7/6	5/2
	Zj	1	2	3	0	t5
	$C_J - Z$	0	0	0	*i	

C_J-Z_J is either zero or negative under all columns, The optimal feasible solution has been obtained

$$\frac{5}{2} \quad \frac{5}{2} \quad \frac{5}{2}$$

$$A \; 'A_z 'A \; ' \; 0$$

$$\boxed{Z_{max} = 15}$$

Problem 8. **Use penalty Method to**

minimize $Z = x_1 + 2x_2 + x_3$

Subject to $x_1 + \frac{x_2}{2} + \frac{x_3}{2} \leq 1$

$$\frac{3}{2}x_1 + 2x_2 + x_3 \geq 8$$

$$x_1, x_2, x_3 \geq 0$$

Solution : 1. Standard form

$$\text{minimize } z = x_1 + 2x_2 + x_3 + 0S_1 + 0S_2$$

$$\text{Subject to } x_1 + \frac{x_2}{2} + \frac{x_3}{2} + S_1 = 1$$

$$\frac{3}{2}x_1 + 2x_2 + x_3 - S_2 + A_1 = 8$$

$$x_1, x_2, x_3, S_1, S_2 \geq 0$$

2. ibfs

Setting $x_1 = x_2 = x_3 = S_2 = 0$

$S_1 = 1, A_1 = 8$

3. Iterations for optimal Solution

	C_J	1	2	1	0	0	M		
C_B	Basis	x_1	x_2	x_3	S_1	S_2	A_1	b	θ
0	S_1	1	1/2	1/2	1	0	0	1	2 ←
M	A_1	3/2	2	1	0	-1	1	8	4
	Z_J	3/2M	2M	M	0	-M	M	8M	
	$C_J - Z_J$	$1 - \frac{3}{2}M$	2 - 2M	1 - M	0	M	0		

↑

	C_J	1	2	1	0	0	M	
C_B	Basis	x_1	x_2	x_3	S_1	S_2	A_1	b
2	x_2	2	1	1	2	0	0	2
M	A_1	-5/2	0	-1	-4	-1	1	4
	Z_J	$4-\frac{5}{2}M$	2	2-M	4-4M	-M	M	4+4M
	C_J-Z_J	$-3+\frac{5}{2}M$	0	-1+M	-4+4M	M	0	

Since C_Z is non negative under all columns so optimality test passed since A1appears in he basis at a positive value, the given problem has no feasible solution

---

Problem 9. **Use the two phase simplex method to**

Maximize $Z = 5x_1 - 4x_2 + 3x_3$

Subject to $2x_1 + x_2 - 6x_3 = 20$

$6x_1 + 5x_2 + 10x_3 \leq 76$

$8x_1 - 3x_2 + 6x_3 \leq 50$

$x_1, x_2, x_3 \geq 0$

Solution : Step 1. Standard form

$$2x_1 + x_2 - 6x_3 + A_1 = 20$$
$$6x_1 + 5x_2 + 10x_3 + S_1 = 76$$
$$8x_1 - 3x_2 + 6x_3 + S_2 = 50$$
$$x_1, x_2, x_3, S_1, S_2, A_1 \geq 0$$

The new objective function

minimize $w = A_1$

Subject to

$$2x_1 + x_2 - 6x_3 + 0S_1 + 0S_2 + A_1 = 20$$
$$6x_1 + 5x_2 + 10x_3 + S_1 + 0S_2 + 0A_1 = 76$$
$$8x_1 - 3x_2 + 6x_3 + 0S_1 + S_2 + 0A_1 = 50$$
$$x_1, x_2, x_3, S_1, S_2, A_1 \geq 0$$

Step 2. ibfs

$$x_1 = x_2 = x_3 = 0$$
$$A_1 = 20$$
$$S_2 = 76$$
$$S_3 = 50$$

Step 3. Iterations for optimal Solution :

	C_J	0	0	0	0	0	1		
C_B	Basis	x_1	x_2	x_3	S_1	S_2	A_1	b	θ
1	A_1	2	1	-6	0	0	1	20	10
0	S_1	6	5	10	1	0	0	76	$\frac{38}{3}$
0	S_2	(8)	-3	6	0	1	0	50	$\frac{25}{4}$ ← Key row
	Z_J	2	1	-6	0	0	1	20	
	C_J-Z_J	-2	-1	6	0	0	0		

	C_J	0	0	0	0	0	1.			
C_B	Basis	x_1	x_2	x_3	S_1	S_2	A_1	b	θ	
1	A_1	0	(7/4)	-15/2	0	-1/4	1	15/2	30/7	← Key row
0	S_2	0	29/4	11/2	1	-3/4	0	77/2	154/29	
0	x_1	1	-3/8	3/4	0	1/8	0	25/4	-50/3	
	Z_J	0	7/4	-15/2	0	-1/4	1	15/2		
	C_J-Z_J	0	-7/4	15/2	0	1/4	0			

	C_J	0	0	0	0	0	1	
C_B	Basis	x_1	x_2	x_3	S_1	S_2	A_1	b
0	x_2	0	1	-30/7	0	-1/7	4/7	30/7
0	S_2	0	0	256/7	1	2/7	-29/7	52/7
0	x_1	1	0	-6/7	0	1/14	3/14	55/7
	Z_J	0	0	0	0	0	0	0
	C_J-Z_J	0	0	0	0	0	1	

Since C-Z is non negative under all columns and no artificial variable appears in the basis.

2nd phase (deleting artificial variable column)

	Cj	5	- 4	3	0	0	
Cq	Busts	xt	*z	'r	St	St	6
- 4	z_2	0	1	- 30/7	0	- 1/Z	H/7
0		0	0	256/7	1	2/7	52/7
5	z	1	0	- 6/7	0	1/14	55/7
	Zj	5	- 4	90/7	0	13/ 14	155/ 7
	Cj - Zj	0	0	- 69/7	0	- 13/14	

Since - Zj is either negative or zero under all columns, optimality test has passed.

55 30

’+' 7 ‘*' 7 ‘*'

Zmax * 155 7

variable.

New values of the Entering Variable = Old values of the leaving variable / Key element.

Thereafter, we should write the new values of the other left out rows. The formula is New values of the Left out row= Old values of the left out row – (New values of the entering variable X value in the key column of the Old left out row)

UNIT – 1 – INTRODUCTION AND LINEAR PROGRAMMING

QUESTION BANK

PART – A

1. Define Operations Research.
2. Write the stages in operations research?
3. What are the areas in which operations research is being applied?
4. Name the models being classified based on nature.
5. What do you mean by simulation model?
6. Enumerate the limitations of operations research.
7. What is meant by Linear programming problem?
8. Write the steps involved in formulation of linear programming problem.
9. What are the decision variables in LPP?
10. A person wants to decide the constituents of a diet which will fulfill his daily requirement of protein, fat, and carbohydrates at minimum cost. The choice is to be made from 4
11. different types of food.

Food type	(Yield/unit)			Cost/unit
	protein	fat	carbohydrates	
1	3	2	6	45
2	4	2	4	40
3	8	7	7	85
4	6	5	4	65
Min	800	200	700	

requirement

12. Consider food stuff A&B. These contain three vitamins V1, V2, V3. Minimum daily requirement of V1 is 1mg, V2 is 50mg and V3 is 10mg. Suppose food A contain 1mg of V1, 100mg of V2 and 10mg of V3. and food B contain 1mg of V1,10mg of V2. Cost of 1 unit of food A is ₹1 and food B ₹1.5.

13. Formulate the LPP. An organization wants to produce Tables & Chairs Profit of 1 table is ₹100 and profit of 1 Chair is ₹30.

	Tables	Chairs	Maximum hours available
Cutting (hours)	4	1	300
Painting (hours)	1	½	100

PART - B

11. Briefly discuss about Models in operations research.

OR

12. Solve the LPP using Graphical method for the given formulation.

MAX Z=28x+30y

Subject to x+3y≤18, 3x+y≤8, 4x+5y≤30 (x, y ≥ 0).

13. Define operations research. Give the scope, characteristics of operations research.

OR

14. Solve the given LPP using Simplex method.

MAX $Z=6X_1+12X_2$

Subject to $3X_1+4X_2\leq 12$, $10X_1+5X_2\leq 20$, $(X_1, X_2 \geq 0)$.

15. Solve the LPP using Graphical method for the given formulation. MAX Z=5x-2y

Subject to 2x+y≤9, 2x-4y48, 3x+2y≤3 (x, y ≥ 0).

OR

16. Solve the given LPP using Simplex method. MAX $Z=6X_1+12X_2$

Subject to $3X_1+4X_2 \leq 12$, $10X_1+5X_2 \leq 20$, $(X_1, X_2 \geq 0)$.

17. Solve the LPP using Graphical method for the given formulation. MAX $Z=4X_1+2X_2$
    Subject to $2X_1+3X_2 \leq 18$, $X_1+X_2 \geq 10$ $(X_1, X_2 \geq 0)$.

OR

18. Solve the given LPP using Simplex method. MINZ= -
    40X1 - 100X2
    Subject to $10X_1+5X_2 \leq 250$, $2X_1+5X_2 \leq 100$, $2X_1+3X_2 \leq 90$ $(X_1, X_2 \geq 0)$.

19. Solve the LPP using Graphical method for the given formulation. MAX $Z=2X_1+4X_2$
    Subject to $X_1+X_2 \leq 14$, $3X_1+2X_2 \geq 30$, $2X_1+X_2 \leq 18$ $(X_1, X_2 \geq 0)$.

OR

20. Solve the LPP using Graphical method for the given formulation
    MAX Z=2X1 + 4X2
    Subject to X1+X2≤14, 3X1+2X2≥30,
    2X1+X2≤18, (X1, X2≥0)

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – IV – PRODUCTION MANAGEMENT – SMTA5205

REPLACEMENT MODEL

If any equipment or machine is used for a long period of time, due to wear and tear, the item tends to worsen. A remedial action to bring the item or equipment to the original level is desired. Then the need for replacement becomes necessary. This may be due physical impairment, due to normal wear and tear, obsolescence etc. The resale value of the item goes on diminishing with the passage of time.

The depreciation of the original equipment is a factor, which is responsible not to favor replacement because the capital is being spread over a long time leading to a lower average cost. Thus there exists an economic trade-off between increasing and decreasing cost functions. We strike a balance between the two opposing costs with the aim of obtaining a minimum cost.

Replacement model aims at identifying the **time** at which the assets must be replaced in order to minimize the cost.

4.1 REASONS FOR REPLACEMENT OF EQUIPMENT:

1. Physical impairment or malfunctioning of various parts refers to

- The physical condition of the equipment itself
- Leads to a decline in the value of service rendered by the equipment
- Increasing operating cost of the equipment
- Increased maintenance cost of the equipment
- Or a combination of the above.

2. Obsolescence of the equipment, caused due to improvement in the existing tools and machinery mainly when the technology becomes advanced.

3. When there is sudden failure or breakdown.

4.2 REPLACEMENT MODELS:

- **Assets that fails Gradually:**

  Certain assets wear and tear as they are used. The efficiency of the assets decline with time. The maintenance cost keeps increasing as the years pass by eg. Machinery, automobiles, etc.

  1. Gradual failure without taking time value of money into consideration
  2. Gradual failure taking time value of money into consideration

- **Assets which fail suddenly**

  Certain assets fail suddenly and have to be replaced from time to time eg. bulbs.

  1. Individual Replacement policy (IRP)
  2. Group Replacement policy (GRP)

4.3.1 Assests that fails Gradually

4.3.1.1 Gradual failure without taking time value of money into consideration

As mentioned earlier the equipments, machineries and vehicles undergo wear and tear with the passage of time. The cost of operation and the maintenance are bound to increase year by year. A stage may be reached that the maintenance cost amounts prohibitively large that it is better and economical to replace the equipment with a new one. We also take into account the salvage value of the items in assessing the appropriate or opportune time to replace the item. We assume

that the details regarding the costs of operation, maintenance and the salvage value of the item are already known

- **Procedure for replacement of an asset that fails gradually (without considering Time value of money):**

a) Note down the years
b) Note down the running cost 'R' (Running cost or operating cost or Maintenance cost or other expenses)
c) Calculate Cumulative the running cost '∑R'
d) Note down the capital cost 'C'
e) Note down the scrap or resale value 'S'
f) Calculate Depreciation = Capital Cost – Resale value
g) Find the Total Cost
   Total Cost = Cumulative Running cost + Depreciation
h) Find the average cost
   Average cost = Total cost/No. of corresponding year

i) Replacement decision: Average cost is minimum (Average cost will decrease and reach minimum, later it will increase)

Year	Running Cost	Cumulative Running Cost	Capital cost	Salvage value Or Resale value	Depn. = Capital cost – salvage value	Total cost= Cumulative running cost + Depreciation	Average annual cost P_n = Total cost / no. of corresponding year
N	R_n	$\sum R_n$	C	S_n	$C - S_n$	$\sum R_n + C - S_n$	$(\sum R_n + C - S_n)/n$
1	2	3	4	5	6 (4-5)	7 (3+6)	8 (7/1)

4.3.1.2 Gradual failure taking time value of money into consideration

In the previous section we did not take the interest for the money invested, the running costs and resale value. If the effect of time value of money is to be taken into

account, the analysis must be based on an equivalent cost. This is done with the present value or present worth analysis.

For example, suppose the interest rate is given as 10% and Rs. 100 today would amount to Rs. 110 after a year's time. In other words the expenditure of Rs. 110 in year's time is equivalent to Rs. 100 today. Likewise one rupee a year from now is equivalent to (1.1)-1 rupees today and one-rupee in '*n*' years from now is equivalent to (1.1)-n rupees today. This quantity (1.1)-n is called the present value or present worth of one rupee spent 'n' years from now.

- **Procedure for replacement of an asset that fails gradually (with considering Time value of money):**

Assumption:

i. Maintenance cost will be calculated at the beginning of the year
ii. Resale value at the end of the year

Procedure:

a) Note down the years
b) Note down the running cost ‘R’ (Running cost or operating cost or Maintenance cost or other expenses)
c) Write the present value factor at the beginning for running cost
d) Calculate present value for Running cost
e) Calculate Cumulative the running cost ‘$\sum$ R’
f) Note down the capital cost ‘C’
g) Note down the scrap or resale value ‘S’
h) Write the present value factor at the end of the year and also calculate present value for salvage or scrap or resale value.
i) Calculate Depreciation = Capital Cost – Resale value

j) Find the Total Cost = Cumulative Running cost + Depreciation
k) Calculate annuity factor (Cumulative present value factor at the beginning)

l) Find the Average cost = Total cost / Annuity

m) Replacement decision: Average cost is minimum (Average cost will decrease and reach minimum, later it will increase)

Year n	R_n	Pv^{n-1}	R_nPv^{n-1}	$\sum R_nPv^{n-1}$	C	S_n	Pv^n	S_nPv^n	$C - S_nPv^n$	$\sum R_nv^{n-1} + C - S_nPv^n$	$\sum PV^n$	W_n
1	2	3	4(2*3)	5	6	7	8	9(7*8)	10	11(5+10)	12	13

4.3.2 ITEMS THAT FAIL COMPLETELY AND SUDDENLY

There is another type of problem where we consider the items that fail completely. The item fails such that the loss is sudden and complete. Common examples are the electric bulbs, transistors and replacement of items, which follow sudden failure mechanism.

4.3.2.1 INDIVIDUAL REPLACEMENT POLICY (IRP):

Under this strategy equipments or facilities break down at various times. Each breakdown can be remedied as it occurs by replacement or repair of the faulty unit.

Examples: Vacuum tubes, transistors

Calculation of Individual Replacement Policy (IRP):

Average life of an item = $\sum_{i-1}^{n} i * Pi$

Pi denotes Probability of failure during that week **i** denotes no. of weeks

No. of failures = $\frac{\text{Total no. of items}}{\text{Average life of an item}}$

Total IRP Cost = No. of failures * IRP cost

4.3.2.2 GROUP REPLACEMENT

As per this strategy, an optimal group replacement period '*P*' is determined and common preventive replacement is carried out as follows.

(a) Replacement an item if it fails before the optimum period '*P*'.

(b) Replace all the items every optimum period of '*P*' irrespective of the life of individual item. Examples: Bulbs, Tubes, and Switches.

Among the three strategies that may be adopted, the third one namely the group replacement policy turns out to be economical if items are supplied cheap when purchased in bulk quantities. With this policy, all items are replaced at certain fixed intervals.

4.3.4.1 Procedure for Group Replacement Policy (GRP):

1. Write down the weeks
2. Write down the individual probability of failure during that week
3. Calculate No. of failures:

   N_0 - No. of items at the beginning

   $^{st}N_1$ - No. of failure during 1 week (N_0P_1)

   N_2 - No. of failure during 2^{nd} week ($N_0P_2 + N_1P_1$)

   $^{rd}N_3$ - No. of failure during 3 week ($N_0P_3 + N_1P_2 + N_2P_1$)
4. Calculate cumulative failures
5. Calculate IRP Cost = Cumulative no. of failures * IRP cost
6. Calculate and write down GRP Cost = Total items * GRP Cost
7. Calculate Total Cost = IRP Cost + GRP Cost
8. Calculate Average cost = Total cost / no. of corresponding year

4.4 GAME THEORY

A competitive situation in business can be treated similar to a **game**. There are two or more players and each player uses a strategy to out play the opponent.

A strategy is an action plan adopted by a player in-order to counter the other player.In our of game theory we have two players namely Player A and Player B.
The basic objective would be that

Player A – plays to **Maximize profit** (offensive) - Maxi (min) criteria

Player B – plays to **Minimize losses** (defensive) - Mini (max) criteria

The Maxi (Min) criteria is that – Maximum profit out of minimum possibilities

The Mini (max) criteria is that – Minimze losses out of maximum possibilities.

Game theory helps in finding out the best course of action for a firm in view of the anticipated counter-moves from the competing organizations.

4.4.1 Characteristics of a game

A competitive situation is a competitive game if the following properties hold good

1. The number of competitors is finite, say N.
2. A finite set of possible courses of action is available to each of the N competitors.

3. A play of the game results when each competitor selects a course of action from the set of courses available to him. In game theory we make an important assumption that all the players select their courses of action simultaneously. As a result no competitor will be in a position to know the choices of his competitors.

4. The outcome of a play consists of the particular courses of action chosen by the individual players. Each outcome leads to a set of payments, one to each player, which may be either positive, or negative, or zero.

4.4.2 TERMINOLOGIES

Zero Sum game because the Gain of A – Loss of B = 0. In other words, the gain of Player A is the Loss of Player B.

Pure strategy If a player knows exactly what the other player is going to do, a deterministic situation is obtained and objective function is to minimize the gain Therefore the pure strategy is a decision rule always to select a particular course of action.

Mixed strategy If a player is guessing as to which activity is to be selected by the other on any particular occasion, a probabilistic situation is obtained and objective function is to maximize the

expected gain. Thus, the mixed strategy is a selection among pure strategies with fixed probabilities.

Optimal strategy The strategy that puts the player in the most preferred position irrespective of the strategy of his opponents is called an optimal strategy Any deviation from this strategy would reduce his payoff.

Saddle Point : If the Maxi (min) of A = Mini (max) of B then it is known as the Saddle Point Saddle point is the number, which is lowest in its row and highest in its column. When minimax value is equal to maximin value , the game is said to have saddle point. It is the cell in the payoff matrix which satisfies minimax to maximin value

Value of the Game : It is the average wining per play over a long no. of plays. It is the expected pay off when all the players adopt their optimum strategies .If the value of game is zero it is said to be a fair game , If the value of game is not zero it is said to be a unfair game . In all problems relating to game theory, first look for saddle point, then check out for rule of dominance and see if you can reduce the matrix.

Rule of Dominance:

The dominance and modified dominance principles and their applications for reducing the size of a game with or without a saddle point. If every value of one strategy of A is lesser than that of the other strategy of A, Then A will play the strategy with greater values and remove the strategy with the lesser payoff values.

If every value of one strategy of B is greater than that of other strategy of B, B will play the lesser value strategy and remove the strategy with higher payoff values.

Dominance rule for the row

If all the elements in a particular row is lower than or equal to all the elements in another row, then the row with the lower items are said to be dominated by row with higher ones, Then the row with lower elements will be eliminated.

Dominance rule for the column

If all the elements in a particular column is higher than or equal to all the elements in another column, then the column with the higher items are said to be dominated by column with lower ones, Then the column with higher elements will be eliminated.

Modified Dominance Rule

In few cases, if the given strategy is inferior to the average of two or more pure strategies, then the inferior strategy is deleted from the pay-off matrix and the size of the matrix is reduced considerably. In other words, if a given row has lower elements than the elements of average of two rows then particular row can be eliminated. Similarly if a given column has higher elements than the elements of average of two columns then particular column can be eliminated. Average row/column cannot be eliminated under any circumstances.. This type of dominance property is known as the modified dominance property

4.4.3 Graphical Method

If one of the players, play only two strategies or if the game can be reduced such that one of the players play only two strategies. Then the game can be solved by the graphical method.

In case the pay-off matrix is of higher order (say m x n), then we try to reduce as much as possible using dominance and modified dominance ,f we get a pay-off matrix of order 2 x n or n x 2 we try to reduce the size of the pay-off matrix to that of order 2 x 2 with the graphical method so that the value of game could be obtained

.

4.4.4 Managerial Applications of the Theory of Games

The techniques of game theory can be effectively applied to various managerial problems as detailed below:

1. Analysis of the market strategies of a business organization in the long run.
2. Evaluation of the responses of the consumers to a new product.
3. Resolving the conflict between two groups in a business organization.
4. Decision making on the techniques to increase market share.
5. Material procurement process.
6. Decision making for transportation problem.
7. Evaluation of the distribution system.
8. Evaluation of the location of the facilities.
9. Examination of new business ventures and
10. Competitive economic environment

GAME THEORY

PURE STRATEGIES

1. Solve the game whose payoff matrix is given below

	B1	B2	B3
A1	-2	5	-3
A2	1	3	5
A3	-3	-7	11

2.

	B1	B2	B3
A1	0	-4	-2
A2	3	-5	1
A3	-2	-1	6
A4	1	0	4

Problems

Problem 1. The cost of a machine is Rs. 6100/- and its scrap The maintenance costs found from experience are as follows:

Year	1	2	3	4	5	6	7	8
Maintenance cost	100	250	400	600	900	1200	1600	2000

When should the machine be replaced ?

Ans. Let it is profitable to replace the machine after n years. The n is determined by the minimum value of T_{avg}.

Years service	Purchase price-scrap value	Annual maintenance cost	Summation of maintenance cost	Total cost	Avg. annual cost (T_{avg})
1.	6000	100	100	6100	6100
2.	6000	250	350	6350	3175
3.	6000	400	750	6750	2250
4.	6000	600	1350	7350	1837.50
5.	6000	900	2250	8250	1650
6.	6000	1200	3450	9450	1575 Min
7.	6000	1600	5050	11050	1578
8.	6000	2000	7050	13050	1631

The avg. annual cost is minimum Rs. should be replaced after 6 years of use.

(1575/-) during the sixth year. Hence the m/c

Problem 2. A machine owner finds from his past records that the costs per year of maintaining a machine whose purchase price is Ks. 6000 are as given below

Year	1	2	3	4	5	6	7	8
Maintenance cost	1000	1200	1400	1800	2300	2800	3400	4000
Cost Resale price	3000	1500	750	375	200	200	200	200

Determine at what age is a replacement due?

Ans. Capital cost C = 6000/-. Let it be profitable to replace the. machine after n

years. Then n should be determined by the minimum value of Tav•

Year of service	Resale value	Purchase Price Resale value	Annual Maintenance cost	Summation of maintenance cost	Total Cost	Average annual cost
1.	3000	3000	1000	1000	4000	4000
2.	1500	4500	1200	2200	6700	3350
3.	750	5250	1400	3600	8850	2950
4.	375	5625	1800	5400	11025	2756.25
5.	200	5800	2300	7700	13500	2700
6.	200	5800	2800	10500	16300	2716.66
7.	200	5800	3400	13900	19700	2814.28
8.	200	5800	3400	17300	23100	2887.5

We observe from the table that avg. annual cost is minimum (Rs. 2700/-). Hence the m/c should replace at the end of 5th year.

Type B. Replacement of items whose maintenance costs increase with time and value of money also changes with time.

The machine should be replaced if the next period's cost is greater than weighted average of previous cost.

Discount rate [Present worth factor (PWF)

$$V = \frac{1}{1+i}$$

$V_n = (V)^{n-1}$

n – no. of year

i – annual interest rate

V_n – PWF of n^{th} year.

Problem 3. A machine costs Rs. 500/— Operation and Maintenance cost are zero for the first year and increase by Rs. 100/— every year. If money. is worth 5% every year, determine the best age at which the machine should be replaced. The resale value of the machine is negligible small. What is the weighted average cost of owning and operating the machine?

Ans. Discount rate $V = \frac{1}{1+i} = \frac{1}{1+0.05} = 0.9524$

Discount rate for I[st] year $V_n = \left(\frac{1}{1+i}\right)^{n-1}$

$V_1 = (0.9524)^0 = 1$

2[nd] year $V_2 = (0.9524)^1 = 0.9524$

3[rd] year $V_3 = (0.9524)^2 = 0.9070$

4[th] year $V_4 = (0.9524)^3 = 0.8638$

5[th] year $V_5 = (0.9524)^4 = 0.8227$

Years of service (n)	Maintenance cost (Rs)	Discount factor $(V)^{n-1}$	Discounted cost	Summation of cost of m/c and maint. Cost	Summation of discount factor	Weighted average cost
1	0	1.0000	0.00	500.00	1.0000	500
2	100	0.9524	95.24	595.24	1.9524	304.88
3	200	0.9070	181.40	776.64	2.8594	217.61 min
4	300	0.8638	259.14	1035.78	3.7232	278.20
5	400	0.8227	329.08	1364.86	4.5459	300.25

M/c su1d be replaced at the end of 3C1 year.

Problem 3. Purchase price of a machine is Rs. 3000/— and its running cost is given in the table below. If should be replaced. the discount rate is 0.90. Find at what age the machine

Year	1	2	3	4	5	6	7
Running cost (Rs.)	500	600	800	1000	1300	1600	2000

Ans. V (Discount rate) = 0.90

Year of service (n)	Running cost (Rs.)	Discount factor $(V)^{n-1}$	Discounted cost	Summation of cost of m/c and maint. cost	Summation of discount factor	Weighted average cost
1	500	1	500	3500	1	3500
2	600	0.90	540	4040	1.9	2126.31
3	800	0.81	648	4688	2.71	1729.88
4	1000	0.729	729	5417	3.439	1575.16
5	1300	0.6561	852.93	6269.93	4.0951	1531.08 min.
6	1600	0.59049	944.78	7214.71	4.6855	1539.79
7	2000	0.5314	1062.8	8277.51	5.2169	1586.6

M/c should be replaced at the end of 5th year.

Problem 4. The following mortality ratio have been observed for a certain type of light bulbs in an installation with 1000 bulbs

End of week	1	2	3	4	5	6
Probability of failure to date	0.09	0.25	0.49	0.85	0.97	1.00

There are a large no. of such bulbs which are to be kept in working order. If a bulb fails in service, it cost Rs. 3 to replace but if all the bulbs all replaced in the same operation it can be done for only Rs. 0.70/— a bulb. It is proposed to replace all bulbs at fixed intervals, whether or not they have burnt out and continue replacing burnt out bulb as they fail.

(a) What is the best interval between group replacement?

(b) Also establish if the policy, as determined by you is superior to the policy of replacing bulbs as and when they, fail, there being nothing like group replacement.

(c) At what group replacement price per bulb, would a policy of strictly individual replacement become preferable to the adopted policy?

Solution : Let p. be the probability that a new light bulbs fail during the 1th wek of the life.

$P_1 = 0.09$

$P_2 = 0.25 - 0.09 = 0.16$

$P_3 = 0.49 - 0.25 = 0.24$

$P_4 = 0.85 - 0.49 = 0.36$

$P_5 = 0.97 - 0.85 = 0.12$

$P_6 = 1.00 - 0.97 = 0.03$

Week	Expected no. of failure (N)	
0	$N_0 = N_0$	
1	$N_1 = 1000 \times 0.09$	= 90
2	$N_2 = 1000 \times 0.16 + 90 \times 0.09$	= 168
3	$N_3 = 1000 \times 0.24 + 90 \times 0.16 + 168 \times 0.09$	= 269
4	$N_4 = 1000 \times 0.36 + 90 \times 0.24 + 168 \times 0.16 + 269 \times 0.09$	= 432
5	$N_5 = 1000 \times 0.12 + 90 \times 0.36 + 168 \times 0.24 + 269 \times 0.16 + 432 \times 0.09$	= 274
6.	$N_6 = 1000 \times 0.03 + 90 \times 0.12 + 168 \times 0.36 + 269 \times 0.24 + 432 \times 0.16 + 274 \times 0.09$	= 260
and so on		

(*a*) Determination of optimum group replacement interval

Week	Total cost of group replacement	Avg cost/week
1.	1000 × 0.70 + 90 × 3 = 970	970.00
2.	1000 × 0.70 +3 (90+168) = 1474	737.00
3.	1000 × 0.70 + 3 (90 + 168+ 269) = 2281	760.33

INVENTORY

Inventory may be defined a stock of goods, commodities or other economic resources that are stored or reserved for smooth and efficient running of business. The inventory may be kept in any one of the following forms:

1. Raw material

2. Work-in progress
3. Finished goods

If an order for a product is receive, we should have sufficient stock of materials required for manufacturing the item in order to avoid delay in production and supply. Also there should not be over stock of materials and goods as it involves storage cost and wastage in storing. Therefore inventory control is essential to promote business. Maintaining inventory helps to run the business smoothly and efficiently and also to provide adequate service to the customer. Inventory control is very useful to reduce the cost of transportation and storage.

A good inventory system, one has to address the following questions quantitatively and

qualitatively.

- What to order?
- When to order?
- How much to order?
- How much to carry in an inventory?

Objectives of inventory management/Significance of inventory management

To maintain continuity in production.

To provide satisfactory service to customers.

To bring administrative simplicity.

To reduce risk.

To eliminate wastage.

To act as a cushion against high rate of usage.

To avoid accumulation of inventory.

To continue production even if there is a break down in few machinery.

To ensure proper execution of policies.

To take advantages of price fluctuations and buy economically.

Costs involved in inventory

1. **Holding Cost (Carrying or Storage Cost)**

   It is the cost associated with the carrying or holding the goods in stock. It includes storage cost, depreciation cost, rent for godown, interest on investment locked up, record keeping and administrative cost, taxes and insurance cost, deterioration cost, etc. It is denoted by 'C'.

2. **Setup Cost/ Ordering Cost**

   Ordering cost is associated with cost of placing orders for procurement of material or finished goods from suppliers. It includes, cost of stationery, postage, telephones, travelling expenses, handling of materials, etc. (Purchase Model)Setup cost is associated with production. It includes, cost involved in setting up machines for production run. (Production Model). Both are denoted by 'S'.

3. Purchase Cost/Production Cost

When the organization purchases materials from other suppliers, the actual price paid for the material will be called the purchase cost.

When the organization produces material in the factory, the cost incurred for production of material is called as production cost. Both are denoted by 'P'.

4. Shortage Cost

If the inventory on hand is not sufficient to meet the demand of materials or finished goods, then it results in shortage of supply. The cost may include loss of reputation, loss of customer, etc.

Total incremental cost = Holding Cost + Setup Cost/ Ordering Cost

Total Cost = Purchase Cost/ Production Cost + Shortage Cost + Total Incremental cost.

Demand can be classified broadly into two categories:

Deterministic i.e., a situation when the demand is known with certainty. And, deterministic demand can either be *static* (where demand remains constant over time) or it could be *dynamic* (where the demand, though known with certainty, may change with time).

Probabilistic (Stochastic) refers to situations when the demand is *random* and is governed by a *probability density function* or *probability mass function.* Probabilistic demand can also be of two types - *stationary*(in which the demand probability density function remains unchanged over time),and *non-stationary,* where the probability densities vary over time.

Deterministic Inventory Models

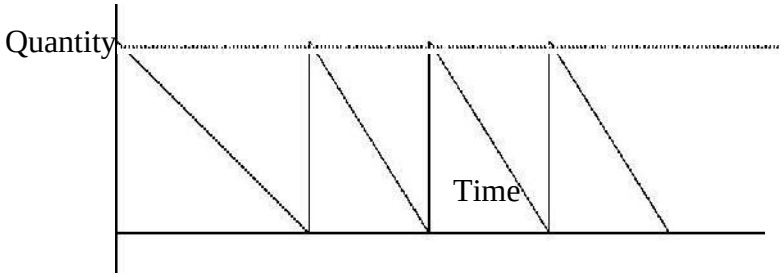
i. Model I: Purchasing model without shortages
ii. Model II: Production model without shortages
iii. Model III: Purchasing model with shortages

iv. Model IV: Production model with shortages

Model I: Purchasing model without shortages

Assumptions

- Demand(D) per year is known and is uniform
- Ordering cost(S) per order remains constant
- Carrying cost(C) per unit remains constant
- Purchase price(P) per unit remains constant
- No Shortages are allowed. As soon as the level of inventory reaches zero, the inventory is replenished back. Lead time is Zero.

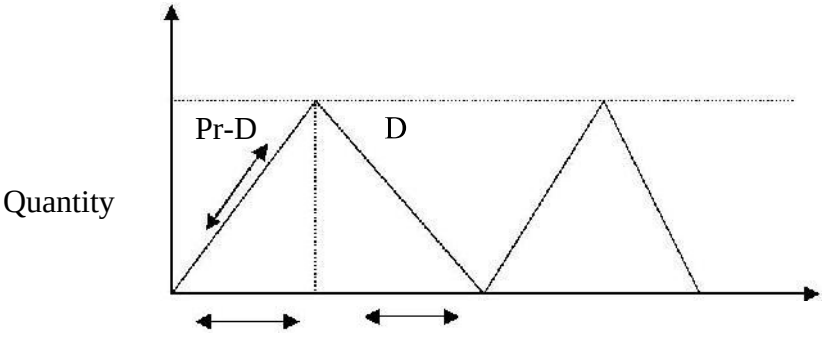


Inventory decreases at the rate of 'D' As soon as the level of inventory reaches zero, the inventory is replenished back

Model II: Production model without shortages

Assumptions

- Demand(D) per year is known and is uniform
- Setup cost (S) per production run remains constant
- Carrying cost(C) per unit remains constant
- Production cost per unit(P) per unit remains constant
- No Shortages are allowed. As soon as the level of inventory reaches zero, the inventory is replenished back.



T1 T2 Pr = Production Rate
D = Demand Rate

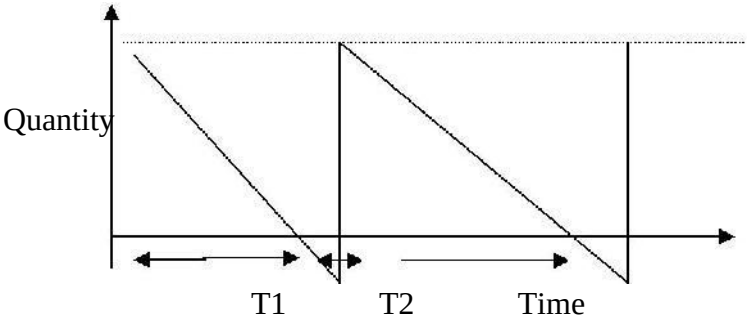
T1 is the time taken when manufacturing takes place at the rate of Pr and demand at the rate of

D. So the stock is built up at the rate of (Pr – D). During t2 there is no production only usage of stock. Hence, stock is decreased at the rate of 'D'. At the end of t2, stock will be nil.

Model III: Purchasing model with shortages

Assumptions

- Demand(D) per year is known and is uniform
- Ordering cost(S) per order remains constant
- Carrying cost(C) per unit remains constant
- Purchase price(P) per unit remains constant
- Shortages are allowed. As soon as the level of inventory reaches zero, the inventory is replenished back with lead time.
- Shortage cost (sh) per unit remains constant



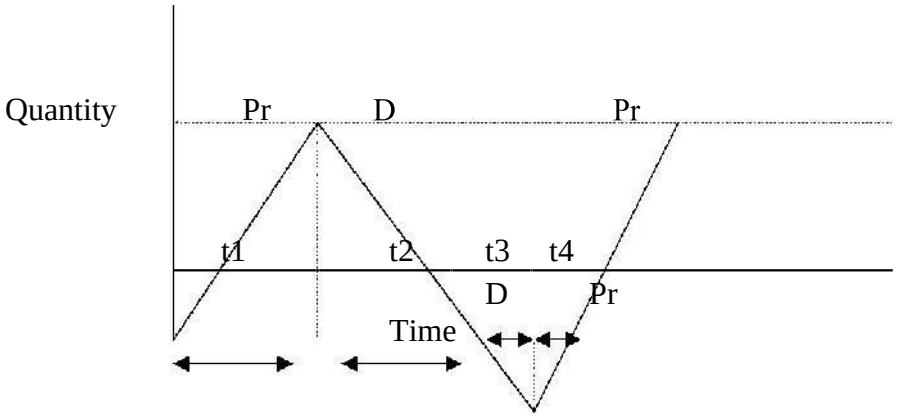
T1 is the time during which stock is nil.During T2 shortage occur and at the end of T2 stock is replenished back.

Model IV: Production model with shortages

Assumptions

Demand(D) per year is known and is uniform

- Setup cost (S) per production run remains constant
- Carrying cost(C) per unit remains constant
- Production cost per unit(P) per unit remains constant
- Shortages are allowed. As soon as the level of inventory reaches zero, the inventory is replenished back with lead time.
- Shortage cost (Sh) per unit remains constant



T1 is the time taken when manufacturing takes place at the rate of Pr and demand at the rate of D. So the stock is built-up at the rate of (Pr – D). During t2 there is no production only usage of stock. Hence, stock is decreased at the rate of 'D'. At the end of t2, stock will be nil. During T3 shortage exists at the rate of 'D'. During T4 production begins stock builds and shortage decreases at the rate of 'Pr-D'

Inventory basic terminologies

- EOQ- Economic order quantity – The optimum order per order quantity for which total inventory cost is minimum.
- EBQ- Economic batch quantity – The optimum manufacturing quantity in one batch for which total inventory cost is minimum.
- Demand Rate – rate at which items are consumed
- Production rate- rate at which items are produced
- Stock replenishment rate
    - Finite rate – the inventory builds up slowly /step by step(production model)
    - Instantaneous rate – rate at which inventory builds up from minimum to maximum instantaneously (purchasing model)
- Lead time- Time taken by supplier to supply goods
- Lead time demand it is the demand for goods in the organization during lead time.
- Reorder level- the level between maximum and minimum inventory at which purchasing or manufacturing activities must start from replenishment.

  Reorder level = Buffer stock+ Lead time demand
- Buffer stock- to face the uncertainties in consumption rate and lead time , an extra stock is

maintained. This is termed as buffer stock:

Buffer stock = (Maximum Lead time – Average Lead time) x Demand per month

- Maximum Inventory Level: Maximum quantity that can be allowed in the stock: Maximum Inventory = EOQ + Buffer stock
- Minimum Inventory Level is the level that is expected to be available when thee supply is due: Minimum Inventory level = Buffer stock
- Average Inventory = (Minimum Inventory + Maximum Inventory)/2
- Order cycle is the period of time between two consecutive placements of orders.

Inventory system followed in a organization:

- Q – System (fixed order quantity system)
- P - System (fixed period system)

Q – System

In a fixed order quantity system means every time an order is placed the quantity order is EOQ.

In Q – System, the period between the orders is not constant:

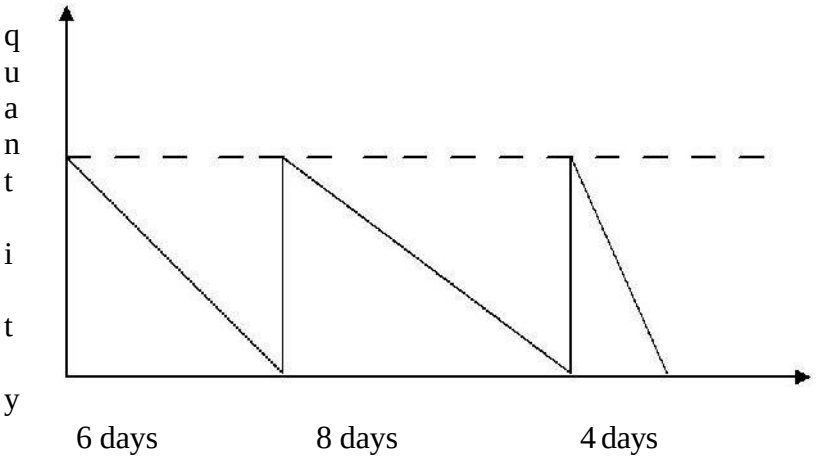
Ex. 1st – 1 month – EOQ

2nd – 1 ½ month – EOQ

3rd - 2 month – EOQ

4th – 15 days - EOQ

Whenever the stock reaches reorder level, next order is placed.



- Reorder level- the level between maximum and minimum inventory at which purchasing or manufacturing activities must start from replenishment.
  Reorder level = Buffer stock+ Lead time demand
- Lead time is the time taken by supplier to supply goods
- Lead time demand it is the demand for goods in the organization during lead time.
- Buffer stock: To face the uncertainties in consumption rate and lead time, an extra stock is maintained. This is termed as buffer stock:
  Buffer stock = (Maximum Lead time – Average Lead time) x Demand per month
- Maximum Inventory Level: Maximum quantity that can be allowed in the stock: Maximum Inventory = EOQ + Buffer stock
- Minimum Inventory Level is the level that is expected to be available when the supply is due:
  Minimum Inventory level = Buffer stock
  Average Inventory = (Minimum Inventory + Maximum Inventory)/2

P – System

Time period between the orders is fixed; hence it is called as Fixed Period System. Period of order is fixed but the quantity will vary. Ex:

1st – 1 month – 1000 units
2nd – 1 month – 1200 units
3rd - 1 month – 950 units

A predetermined level of inventory is fixed and thee order quantity is determined by deducting the level of stock at the time review from P determine level of inventory.

Order quantity = Predetermined level of inventory – level of stock at the time of review

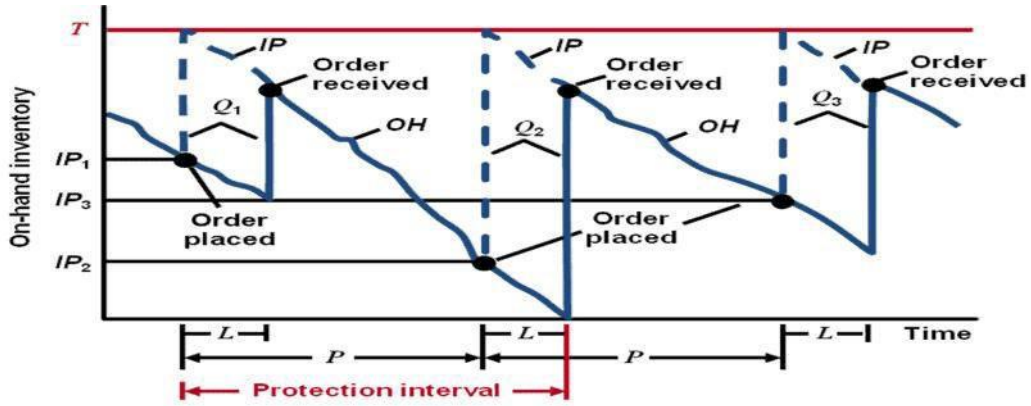


Figure 12.10 – *P* System When Demand Is Uncertain

5.3Inventory Selective Control Techniques

Every organization consumes several items of store. Since all the items are not of equal importance, a high degree control on inventories of each item is neither applicable nor useful. So it becomes necessary to classify items in group depending upon their utility importance. Such type of classification is name as the principle of selective control.

ABC Analysis (Always Better Control)

- A - High value items
- B - Moderate value items
- C - Low value items

ABC analysis is one of the methods for classification of materials. It is based on Parelo's law that a few high usage value items constitute a major part of the inventory while a large bulk of items constitute to very low usage value.

5.4.1.1 PROCEDURE FOR ABC ANALYSIS:

1. Note down the material code.
2. Note down the annual usage in terms of units.
3. Note down the price per unit.
4. Calculate the Annual usage value.

Annual usage value = Quantity used x Price per unit

5. Arrange the materials according to the value in descending order.
6. Find out the percentage contribution of each material to the total value.
7. Find out the percentage contribution of each material towards the total quantity.

8. Cumulate the % contribution towards value.
9. The classification is as follows.

A = 80% contribution B = 15% contribution C = 5% contribution.

5.4.1.2 SIGNIFICANCE OF ABC ANALYSIS

ABC analysis is a very useful technique to classify the materials.

- The control procedure is based on which category the item belongs to. A = Tight control
  B = Moderate control
  C = Very little control.

- The inventory to be maintained is again based on the category A = Low Inventory
  B = Moderate Inventory C = High Inventory.

- The number of suppliers is also based on the category to which it belongs. A = Many suppliers
  B = Moderate No. of suppliers C = Few suppliers.

VEDAnalysis

- V Vital items
- E Essential items
- D Desirable or Durable it ems

HML Analysis

- High price items
- Moderate price items
- Low price items

FNSD Analysis

- F Fast Moving items
- N Normal Moving items
- S Slow Moving items

- D Dead items

Probabilistic Inventory Model

One such model is fixed order quantity model (FOQ).

In this model,

1. The demand (D) is uncertain, you can estimate the demand through any one of the forecasting techniques an d the probability of demand distribution is known.

2. Lead time (L) is uncertain, probability of lead time distribution is known.
3. Cost(C) all the costs are known.
    a. –Inventory holdin g costs C1

    b. –shortage cost C2
4. The optimum order level Z is determined by the following relationship

$$\sum_{d=0}^{z-1} p(d) < \frac{C2}{C1+C2} < \sum_{d=0}^{z} p(d)$$

Stock out Cost/Shortage cost

It is difficult to calculate stock out cost because it consists of components difficult to quantify so indirect way of handling stock out cost is through service levels. Service levels means ability of organization to meet the requirements of the customer as on when he demands for the product. It is measured in terms of percentage.

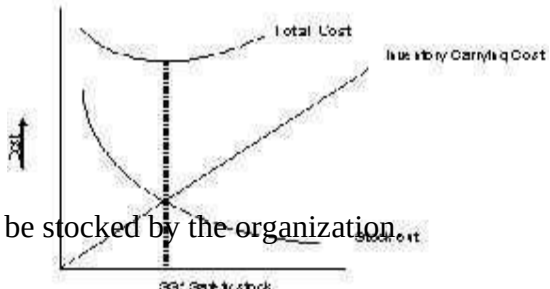
For example: if an organization maintains 90% service level, this means that 10% is “stock out” level. This way the stock out level is addressed.

Safety stock

It is the extra stock or buffer stock or minimum stock. This is kept to take care of fluctuations in

demand and lead time.

If you maintain more safety stock, this helps in reducing the chances of being “stock out”. But at the same time it increases the inventory carrying cost. Suppose the organization maintains less service level that results in more stock out cost but less inventory carrying cost. It requires a tradeoff between inventory carrying cost and stock out cost. This is explained through following Fig.



Safety stock (S.S*) is to be stocked by the organization.

Working of fixed order quantity model

Fixed order quantity system is also known as continuous review system or perpetual inventory system or Q system.

In this system, the ordering quantity is constant. Time interval between the orders is the variable. The system is said to be defined only when if the ordering quantity and time interval between the orders are specified. EOQ provides answer for ordering quantity. Reorder level provides answers for time between orders.

The working and the fixed order quantity model is shown in the below Fig

Application of Fixed Order Quantity System

1. It requires continuous monitoring of stock to know when the reorder point is reached.
2. This system could be recommended to" A" class because they are high consumption items. So we need to have fewer inventories. This system helps in keeping less inventory comparing to other inventory systems.

Advantages:

1. Since the ordering quantity is EOQ, comparatively it is meaningful. You need to have less safety stock. This model relatively insensitive to the forecast and the parameter changes.
2. Fast moving items get more attention because of more usage.

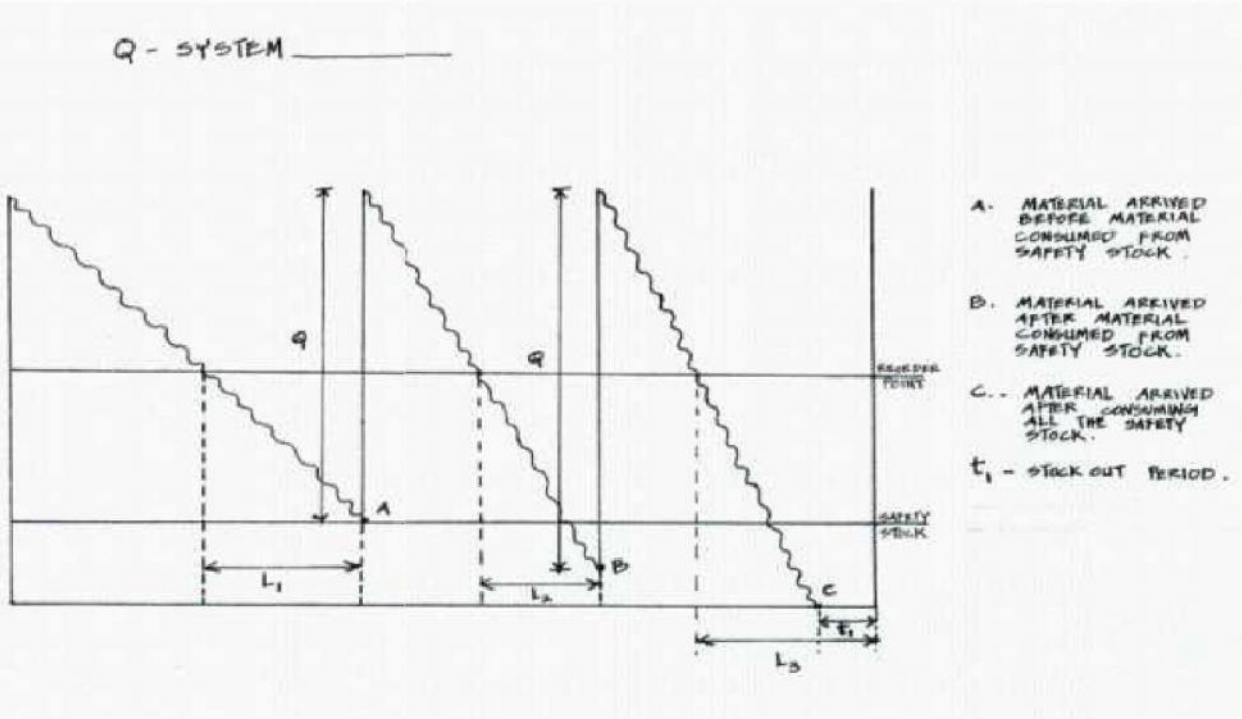
Weakness:

3. We can't club the order for items which are to be procured from one supplier to reduce the ordering cost.

4. There is more chance for high ordering cost and high transaction cost for the items, which follow different reorder level.

5. You can not avail supplier discount. While the reorder level fall in different time periods.

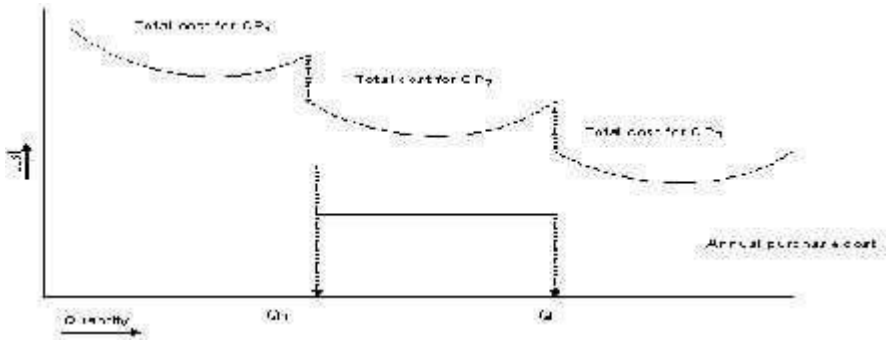
Figure –Fixed Order Quantity Model



QUANTITY DISCOUNT MODEL

As it is mentioned already, the purchase cost becomes relevant with respect to the quantity of order only when the supplier offers discounts. Discounts means if the ordering quantity exceeds particular limit supplier offers the quantity at lesser price per unit. This is possible because the supplier produces more quantity. He could achieve the economy of scale the benefit achieved through economy of scale that he wants to pass it onto customer. This results in lesser price per unit if customer orders more quantity.

If you look at in terms of the customer's perspective customer has also to see that whether it is advisable to avail the discount offered, this is done through a trade off between his carrying inventory by the result of acquiring more quantity and the benefit achieved through purchase price. Suppose if the supplier offers discount schedule as follows,



If the ordering quantity is less than or equal to Q1 then purchase price is Cp1.

If the ordering quantity is more than Q1 and less than Q2 then purchase price is Cp2.

If the ordering quantity is greater than or equal to Q2 then purchase price is Cp3.

Then the curve you get cannot be a continuous total cost curve, because the annual purchase

cost

breaks at two places namely at Q1 and Q2.

STEPS TO FIND THE QUANTITY TO BE ORDERED

1. Find out EOQ for the all price break events. Start with lowest price
2. Find the feasible EOQ from the EOQ's we listed in step 1.
3. Find the total annual inventory cost using the formulae for feasible EOQ =√[2DSC]+D*P
4. Find the total annual inventory cost for the quantity at which price break took place using the following formula.

Total annual inventory cost = TC= (D/Q)*S + (Q/2)*C + D*P

5. Compare the calculated cost in steps 3 and 4. Choose the particular quantity as ordered Quantity at which the total annual inventory cost is minimum.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – V – CONSTRUCTIVE ASSIGNMENTS – SMTA5205

MEASURES OF CENTRAL TENDENCY : CONCEPTS AND FORMULAE

Mean

The mean (also know as average), is obtained by dividing the sum of observed values by the number of observations, *n.*

That is Mean = $\frac{Sum\ of\ all\ observations}{Number\ of\ Observations}$

Median

The median is the middle value of a set of data containing an odd number of values, or the average of the two middle values of a set of data with an even number of values.

Arrange all of the values from lowest to highest. If there are an odd number of entries, the median is the middle value. If there are an even number of entries, the median is the mean of the two middle entries.

Mode

The mode is the most frequently occurring value in the data set. In a data set where each value occurs exactly once, there is no mode.

INDIVIDUAL SERIES	**DISCRETE SERIES**	**CONTINUOUS SERIES**
ARITHMETIC MEAN: Direct Method $\bar{X} = \frac{\sum X}{N}$ Short-cut Method $\bar{X} = A + \frac{\sum d}{N}$ Step-Deviation Method $\bar{X} = A + \frac{\sum d}{N} \times i$	Direct Method $\bar{X} = \frac{\sum f X}{N}$ Short-cut Method $\bar{X} = A + \frac{\sum f d}{N}$ Step-Deviation Method $\bar{X} = A + \frac{\sum f d}{N} \times i$	Direct Method $\bar{X} = \frac{\sum f X}{N}$ Short-cut Method $\bar{X} = A + \frac{\sum f d}{N}$ Step-Deviation Method $\bar{X} = A + \frac{\sum f d}{N} \times i$
MEDIAN: Size of $\left(\frac{N+1}{2}\right)^{th}$ term	Size of $\left(\frac{N+1}{2}\right)^{th}$ term	Size of $\left(\frac{N}{2}\right)^{th}$ term Median = $L + \frac{N/2 - c.f.}{N} \times i$
MODE: Either by inspection or the value that occurs largest number of times	Grouping Method determines that value around which most of the frequencies are concentrated.	Mode = $L + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times i$

EMPIRICAL RELATION: Mode = 3 Mean – 2 Median		

EXAMPLES

1. Find the mode, median and mean for this set of numbers: 3, 6, 9, 14, 3

Solution

First arrange the numbers from least to greatest: 3, 3, 6, 9, 14

Mode (number seen most often) = 3

3, 3, 6, 9, 14

Median (number exactly in the middle) = 6

3, 3, 6, 9, 14

Mean (add up all the numbers then divide by the amount of numbers) = 7

3 + 3 + 6 + 9 + 14 = 35 35 / 5 = 7

2. Find the mode, median and mean for this set of numbers: 1, 8, 23, 7, 2, 5

Solution

First arrange the numbers from least to greatest: 1, 2, 5, 7, 8, 23

Mode = no mode

Median = 6 (5 + 7 = 12 ; 12 / 2 = 6)

Range = 22 (23 – 1 = 22)

Mean = 7 4/6 (1 + 2 + 5 + 7 + 8 + 23 = 46 ; 46 / 6 = 7 4/6 or 7 2/3. This answer may also be stated in decimals.)

3. From the following data compute Arithmetic Mean

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 – 60
No. of students	5	10	25	30	20	10

Solution:

Marks	Midvalue X	No. of students f	f x
0 – 10	5	5	25
10 – 20	15	10	150
20 – 30	25	25	625
30 – 40	35	30	1050
40 – 50	45	20	900
50 – 60	55	10	550
		N=100	3300

Arithmetic Mean $\bar{X} = \frac{\sum f X}{N} = \frac{3300}{100} = 33$

4. Calculate Arithmetic Mean from the following data

Marks	0 - 10	10 - 30	30 - 60	60 - 100
No. of students	5	12	25	8

Solution:

The class intervals are unequal but still to simplify calculations we can take 5 as common factor.

Marks	Midvalue x	No. of students f	d (x - 45) / 5	f d
0 – 10	5	5	- 8	- 40
10 – 30	15	12	- 5	- 60
30 – 60	25	25	0	0
60 - 100	35	8	7	56
		N= 50		- 44

Arithmetic Mean $\bar{X} = A + \frac{\sum f d}{N} \times i$

A = 45, $\sum f d$ = - 44, N = 50, I = 5

$$\bar{X} = 45 - \frac{44}{50} \times 5 = 45 - 4.4 = 40.6$$

5. Find the missing frequency from the following data

Marks	0 – 10	10 – 20	20 - 30	30 - 40	40 -50	50 – 60
No. of Students	5	15	20	-	20	10

The arithmetic mean is 34 marks.

Solution:

Let the missing frequency be denoted by X

Marks	Midvalue x	f`	f x
0 – 10	5	5	25
10 – 20	15	15	225
20 – 30	25	20	500
30 – 40	35	X	35X
40 -50	45	20	900
50 – 60	55	10	550
		N = 70 + X	2200+35X

$$\bar{X} = \frac{\sum f x}{N} \qquad 34 = \frac{2200 + 35X}{70 + X}$$

34 (70 + X) = 2200 + 35X

$$2380 + 34X = 2200 + 35X$$

$$35X - 34X = 2380 - 2200$$

$$X = 180$$

6. Calculate the Median and Mode from the following data

Central size	15	25	35	45	55	65	75	85
Frequencies	5	9	13	21	20	15	8	3

Solution:

Since we are given central values first we determine the lower and upper limits of the classes. The class interval is 10 and hence the first class would be 10 – 20.

Class	Midvalue X	f	d (x – 55) / 10	f d	c.f.
10 - 20	15	5	- 4	- 20	5
20 - 30	25	9	- 3	- 27	14
30 - 40	35	13	- 2	- 26	27
40 -50	45	21	- 1	- 21	48
50 - 60	55	20	0	0	68
60 - 70	65	15	1	15	83
70 - 80	75	8	2	16	91
80 - 90	85	3	3	9	94
				Σfd = - 54	

Calculation of Median:

Med = size of $\frac{N}{2}$ th term = $\frac{94}{2} = 47$

Median lies in the class 40 – 50

Median = $L + \frac{N/2 - c.f.}{f} \times i$

$$M = 40 + \frac{47 - 27}{21} \times 10 = 40 + 9.524 = 49.524$$

7. Calculate the median and mode of the data given below. Using then find arithmetic mean

Marks	0 – 10	10 – 20	20 - 30	30 - 40	40 -50	50 – 60
No. of Students	8	23	45	65	75	80

Solution:

Marks	f`	c.f.
0 – 10	8	8
10 – 20	15	23
20 – 30	22	45

30 – 40	20	65
40 – 50	10	75
50 – 60	5	80
	N = 80	

Calculation of Median: Med = size of $\frac{N}{2}$ th term = $\frac{80}{2} = 40$ th item

Median lies in the class 20 – 30

$$\text{Median} = L + \frac{N/2 - c.f.}{f} \times i$$

$$M = 20 + \frac{40 - 23}{22} \times 10 = 20 + 7.73 = 27.73$$

Mode lies in the class is 20 – 30

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times i = 20 + \left(\frac{22 - 15}{44 - (15 + 20)} \right) \times 10 = 27.78$$

MEASURES OF DISPERSION: CONCEPTS AND FORMULAE

Standard deviation

Standard deviation measures the variation or dispersion exists from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread over a large range of values.

Standard Deviation = $\sqrt{\frac{\text{Sum of (value of entry - mean of data set)}^2}{\text{Number of Entries}}}$

INDIVIDUAL OBERSERVATIONS	**DISCRETE& CONTINUOUS SERIES**
QUARTILE DEVIATION: Q.D. = $\frac{Q_3 - Q_1}{2}$ Coefficient	Quartile Deviation: Q.D. = $\frac{Q_3 - Q_1}{2}$ Coefficient of Q.D. =

of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$	$\frac{Q_3 - Q_1}{Q_3 + Q_1}$
STANDARD DEVIATION: Actual Mean Method: $\sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$ Assumed Mean Method: $\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$ Step Deviation Method $\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} \times i$ $C.V. = \frac{\sigma}{\bar{X}} \times 100$	Actual Mean Method: $\sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$ Assumed Mean Method: $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$ Step Deviation Method $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$ $C.V. = \frac{\sigma}{\bar{X}} \times 100$

EXAMPLES:

1. Find the Mean and standard deviation from the following distribution

Mid value	12.0	12.5	13.0	13.5	14	14.5	15	15.5	16
No. of Students	2	16	36	60	76	37	18	3	2

Solution:

Midvalue x	No. of Students f	d (x – 14) / 0.5	f d	f d^2
12.0	2	-4	- 8	32
12.5	16	-3	- 48	144
13	36	-2	- 72	144
13.5	60	-1	- 60	60
14	76	0	0	0

14.5	37	1	37	37
15	18	2	36	72
15.5	3	3	9	27
16.0	2	4	8	32
	N= 250		Σfd = - 98	$\Sigma fd^2 = 548$

Mean $\bar{X} = A + \frac{\sum f\ d}{N} \times i = 14 - \frac{98}{250} \times 0.5 = 13.8$

Standard deviation $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$

$$\sigma = \sqrt{\frac{548}{250} - \left(\frac{-98}{250}\right)^2} \times .05 \ = \ 0.715$$

2. Find the Standard deviation and Coefficient of Variation from the following data

Marks	No. of students
Up to 10	12
Up to 20	30
Up to 30	65
Up to 40	107
Up to 50	157
Up to 60	202
Up to 70	222
Up to 80	230

Solution:

Class	Midvalue X	No. of Students f	d (x – 35) / 10	f d	f d^2
0 – 10	5	12	-3	- 36	108
10 - 20	15	18	-2	- 36	72
20 - 30	25	35	-1	- 35	35
30 - 40	35	42	0	0	0
40 -50	45	50	1	50	50
50 - 60	55	45	2	90	180
60 - 70	65	20	3	60	180
70 - 80	75	8	4	32	128
		N= 230		Σfd =125	$\Sigma fd^2 = 753$

Mean $\bar{X} = A + \frac{\sum f\ d}{N} \times i = 35 + \frac{125}{230} \times 10 = \ 40.43$

Standard deviation $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$

$$\sigma = \sqrt{\frac{753}{230} - \left(\frac{125}{230}\right)^2} \times 10 \ = \ 17.26$$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100 \;=\; \frac{17.26}{40.43} \times 100 = 42.69$$

3. The scores of two batsmen A and B in ten innings during a certain season are:

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Find which of the two batsmen more consistent in scoring

Solution:

X	$X - \bar{X}$	$(X - \bar{X})^2$	Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
32	-14	196	19	- 31	961
28	-18	324	31	- 19	361
47	1	1	48	- 2	4
63	17	289	53	3	9
71	25	625	67	17	289
39	-7	49	90	40	1600
10	- 36	1296	10	- 40	1600
60	14	196	62	12	144
96	50	2500	40	-10	100
14	- 32	1024	80	30	900
$\sum X = 460$		$\sum (X - \bar{X})^2 = 6500$	$\sum Y = 500$		$\sum (Y - \bar{Y})^2 = 5968$

Batsman A:

Mean $\bar{X} = \frac{\sum X}{N} = \frac{460}{10} = 46$

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{6500}{10}} = 25.495$$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100 \;=\; \frac{25.495}{46} \times 100 = \;\; 55.42$$

Batsman B:

Mean $\bar{Y} = \frac{\sum Y}{N} = \frac{500}{10} = 50$

$$\sigma = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\frac{5968}{10}} = \;\; 24.43$$

$$C.V. = \frac{\sigma}{\bar{Y}} \times 100 \;=\; \frac{24.43}{50} \times 100 = \;\; 48.86$$

Since Coefficient of Variation is less in the case of Batsman B, we conclude that the Batsman B is more consistent.

4. Calculate the Quartile deviation and the coefficient of quartile deviation from the following data

Marks	No. of students
Below 20	8
Below 40	20
Below 60	50
Below 80	70
Below 100	80

Solution:

Marks	f`	c.f.
0 - 20	8	8
20 - 40	12	20
40 - 60	30	50
60 - 80	20	70
80 -100	10	80
	N= 80	

Q_1 is the size of $N / 4^{th}$ item.
Q_1 lies in the class 20 – 40

$$Q_1 = L + \frac{N/4 - c.f.}{f} \times i = 20 + \frac{20-8}{12} \times 20 = 40$$

Q_3 is the size of $3N / 4^{th}$ item.
Q_3 lies in the class 60 – 80

$$Q_3 = L + \frac{3N/4 - c.f.}{f} \times i = 60 + \frac{60-50}{20} \times 20 = 70$$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{70-40}{2} = 15$$

$$Coefficient\ \ of\ \ QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{30}{110} = 0.273$$

5. Calculate the Inter-Quartile range and the coefficient of quartile deviation from the following data

Marks	No. of students
Above 0	150
Above 10	140
Above 20	100
Above 30	80
Above 40	80

Above 50	70
Above 60	30
Above 70	14
Above 80	0

Solution:

Marks	f`	c.f.
0 - 10	10	10
10 - 20	40	50
20 - 30	20	70
30 - 40	0	70
40 -50	10	80
50-60	40	120
60-70	16	136
70-80	14	150
	N = 150	

Q_1 is the size of $N / 4^{th}$ item. Q_1 lies in the class 10 – 20

$$Q_1 = L + \frac{N/4 - c.f.}{f} \times i = 10 + \frac{37.5 - 10}{40} \times 10 = 16.875$$

Q_3 is the size of $3N / 4^{th}$ item. Q_3 lies in the class 50 – 60

$$Q_3 = L + \frac{3N/4 - c.f.}{f} \times i = 50 + \frac{112.5 - 80}{40} \times 10 = 58.25$$

$$Inter\ Quartile\ Range = Q_3 - Q_1 = 41.375$$

$$Coefficient\ of\ QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{41.375}{75} = 0.55$$

<u>MOMENTS</u>: FORMULAE

Moments about mean $\mu_1 = \frac{\sum(X - \overline{X})}{N} = 0$ $\mu_2 = \frac{\sum(X - \overline{X})^2}{N}$	 $\mu_3 = \frac{\sum(X - \overline{X})^3}{N}$ $\mu_4 = \frac{\sum(X - \overline{X})^4}{N}$
In a Frequency distributiom $\mu_1 = \frac{\sum f\,(X - \overline{X})}{N} = 0$	 $\mu_3 = \frac{\sum f\,(X - \overline{X})^3}{N}$

$\mu_2 = \frac{\sum f (X - \overline{X})^2}{N}$	$\mu_4 = \frac{\sum f (X - \overline{X})^4}{N}$
Moments about arbitrary origin $\mu_1' = \frac{\sum (X - A)}{N}$ $\mu_2' = \frac{\sum (X - A)^2}{N}$	$\mu_3' = \frac{\sum (X - A)^3}{N}$ $\mu_4' = \frac{\sum (X - A)^4}{N}$
In a frequency distribution $\mu_1' = \frac{\sum f (X - A)}{N}$ $\mu_2' = \frac{\sum f (X - A)^2}{N}$	$\mu_3' = \frac{\sum f (X - A)^3}{N}$ $\mu_4' = \frac{\sum f (X - A)^4}{N}$
Moments about mean $\mu_2 = \mu_2' - \mu_1'^2$, $\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3$	$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

SKEWNESS AND KURTOSIS

Karl Pearson's Skewness = Mean - Mode
Bowley's Skewness = $Q_3 + Q_1 - 2$ Med
Karl Pearson's coefficient of Skewness = $\frac{Mean - Mode}{\sigma}$
Bowley's coefficient of Skewness = $\frac{Q_3 + Q_1 - 2Med}{Q_3 - Q_1}$
$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \beta_2 = \frac{\mu_4}{\mu_2^2}$
$\gamma_1 = \frac{\mu_3}{\mu_3^{\frac{3}{2}}}, \gamma_2 = \beta_2 - 3$

1. Calculate the coefficient of skewness by Karl Pearson's method and the values of β_1 and β_2 from the following data

Profits (in lakhs)	No. of companies
10 - 20	18
20 - 30	20
30 - 40	30
40 – 50	22
50 – 60	10

Solution :

Class	Midvalue x	No. of Students F	d (x – 35) / 10	f d	$f d^2$	$f d^3$	$f d^4$
10 - 20	15	18	-2	- 36	72	-144	288
20 - 30	25	20	-1	- 20	35	- 20	20
30 - 40	35	30	0	0	0	0	0
40 -50	45	22	1	22	50	22	22
50 - 60	55	10	2	20	180	80	160
		N= 100		$\Sigma fd = -14$	$\Sigma fd^2 = 154$	$\Sigma fd^3 = -62$	$\Sigma fd^4 = 490$

$$\overline{X} = A + \frac{\Sigma f\, d}{N} \times i = 35 - \frac{14}{100} \times 10 = 33.6$$

Modal class 30 -40

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times i = 30 + \frac{30 - 20}{60 - (20 + 22)} = 35.56$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{154}{100} - \left(\frac{-14}{100}\right)^2} \times 10 = 12.33$$

$$\text{Karl Pearson's coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$= \frac{33.6 - 35.56}{12.33} = -0.159$$

$$\mu_1' = \frac{\Sigma fd}{N} = -0.14$$

$$\mu_2' = \frac{\Sigma f\, d^2}{N} = 1.54$$

$$\mu_3' = \frac{\Sigma f\, d^3}{N} = -0.62$$

$$\mu_4' = \frac{\sum f\, d^4}{N} = 4.9$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 1.5204,$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = 0.0213$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 = 4.735$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0.00045}{3.51458} = 0.000128$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{4.735}{2.312} = 2.048$$

2. By measuring the quartiles find a measure of skewness for the following distribution

Annual Sales	No. of firms
Less than 20	30
Less than 30	225
Less than 40	465
Less than 50	580
Less than 60	634
Less than 70	644
Less than 80	650
Less than 90	665
Less than 100	680

Solution:

Sales	f`	c.f.
10 - 20	30	30
20 - 30	195	225
30 - 40	240	465
40 - 50	115	580
50 - 60	54	634
60 - 70	10	644
70 - 80	6	650
80 - 90	15	665
90 - 100	15	680
	N = 680	

Q_1 lies in the class 20-30

$$Q_1 = L + \frac{N/4 - c.f.}{f} \times i = 20 + \frac{170 - 30}{195} \times 10 = 27.18$$

. Q_3 lies in the class 40 – 50

$$Q_3 = L + \frac{3N/4 - c.f.}{f} \times i = 40 + \frac{510 - 465}{115} \times 10 = 43.9$$

$$\text{Inter Quartile Range} = Q_3 - Q_1 = 41.375$$

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{41.375}{75} = 0.55$$

Median class 30 - 40

$$\text{Median} = L + \frac{N/2 - c.f.}{f} \times i$$

$$= 30 + \frac{340 - 225}{240} \times 10 = 34.79$$

$$\text{Bowley's coefficient of Skewness} = \frac{Q_3 + Q_1 - 2\,\text{Med}}{Q_3 - Q_1}$$

$$= \frac{43.9 + 27.18 - 2(34.79)}{43.9 - 27.18} = 0.09$$

3. Calculate the first four moments about the mean from the following data and also calculate the values of β_1 and β_2

Marks	No. of students
0 - 10	5
10 - 20	12
20 - 30	18
30 - 40	40
40 - 50	15
50 - 60	7
60 - 70	3

Solution :

Class	Midvalue x	No. of Students F	d (x – 35) / 10	f d	f d^2	f d^3	f d^4
0 - 10	5	5	-3	-15	45	-135	405
10 - 20	15	12	-2	-24	48	- 96	192
20 - 30	25	18	-1	-18	18	- 18	18
30 - 40	35	40	0	0	0	0	0
40 - 50	45	15	1	15	15	15	15
50 - 60	55	7	2	14	28	56	112
60 - 70	65	3	3	9	27	81	243
		N= 100		$\Sigma f d = 19$	$\Sigma f d^2 = 181$	$\Sigma f d^3 = 97$	$\Sigma f d^4 = 985$

$$\mu_1' = \frac{\sum fd}{N} \times i = -1.9$$

$$\mu_2' = \frac{\sum f\,d^2}{N} \times i^2 = 181$$

$$\mu_3' = \frac{\sum f\,d^3}{N} \times i^3 = -970$$

$$\mu_4' = \frac{\sum f\,d^4}{N} \times i^4 = 98500$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 177.39,$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = 47.982$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'\left(\mu_1'\right)^2 - 3\left(\mu_1'\right)^4 = 95009.364$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{2302.27}{5581968.75} = 0.0004$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{95009.364}{31467.212} = 3.02$$

4. The first four moments of a distribution of a distribution about x = 2 are -2, 12, -20 and 100. Calculate the moment about mean. Also calculate β_2 and find whether the distribution is leptokurtic or platykurtic.

Solution:

$\mu_1' = -2$, $\mu_2' = 12$, $\mu_3' = -20$, $\mu_4' = 100$

$$\mu_2 = \mu_2' - \mu_1'^2 = 8,$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = 36$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'\left(\mu_1'\right)^2 - 3\left(\mu_1'\right)^4 = 20$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} == 0.3125$$

Since β_2 is less than 3 the distribution is platykurtic.

EXERCISE PROBLEMS:

1. Define measure of central tendency.
2. State Karl Pearson's coefficient of skewness.
3. Find the mode of: 4, 8, 3, 8, 8, 9, 1, 8, 3.
4. Define Kurtosis and write the measures of kurtosis.

5. Compute the median for the following frequency distribution

Class: 0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
F : 32	65	100	184	288	167	98	46	20

6. For a group of 200 candidates, the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the scores were found to be 43 and 35 were missed as 34 and 53 respectively. Find the correlated mean and standard deviation corresponding to the corrected figures.

7. Calculate the Arithmetic M can of the following frequency distribution:

X	0-10	10-20	20-30	30-40	40-50	50-60
F	12	18	27	20	17	6

8. In ten cricket matches two batsmen A and B scored as follows

A	12	115	6	73	7	19	119	36	84	29
B	47	12	16	42	4	51	37	48	13	0

Who is better scorer and who is more consistent

9. Calculate the coefficient of skewness and kurtosis on the moments for the following distribution

x	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
f	1	5	12	22	17	9	4	3	1	1

Karl Pearson Coefficient of Correlation

As a measure of intensity or degree of linear relationship between two variables, Karl Pearson developed a formula called Correlation coefficient (also called as product moment correlation coefficient).

Correlation coefficient between two random variables X and Y usually denoted by r(X,Y) or simply r, is a numerical measure of linear relationship between them and is defined as

$$r(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

$$r(X,Y) = \frac{\dfrac{\sum_{i=1}^{n} x_i y_i}{n} - (\bar{x})(\bar{y})}{\sqrt{\dfrac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2}\ \sqrt{\dfrac{\sum_{i=1}^{n} y_i^2}{n} - (\bar{y})^2}}$$

$$r(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 . \sum_{i=1}^{n} (y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

The correlation coefficient is a dimensionless number; it has no units of measurement. The maximum value r can achieve is 1, and its minimum value is –1. Therefore, for any given set of observations, $-1 \le r \le 1$.

EXAMPLES:

1. Calculate the correlation coefficient between X and Y from the following data:

$$\sum_{i=1}^{15} (X_i - \bar{X})^2 = 136 \quad \sum_{i=1}^{15} (Y_i - \bar{Y})^2 = 138 \qquad \sum_{i=1}^{15} (X_i - \bar{X})(Y_i - \bar{Y}) = 122$$

Solution:

We have $$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} = \frac{122}{\sqrt{136}\sqrt{138}}$$ r(X,Y) = 0.89

Example 2. Some health researchers have reported an inverse relationship between central nervous system malformations and the hardness of the related water supplies. Suppose the data were collected on a sample of 9 geographic areas with the following results:

C.N.S. malformation rate (per 1000 births)	9	8	5	1	4	2	3	6	7
Water hardness(ppm)	120	130	90	150	160	100	140	80	200

Calculate the Correlation Coefficient between the C.N.S. malformation rate and Water hardness.

Solution:

Let us denote the C.N.S. malformation rate by x and water hardness by y. The mean of the x series $\bar{x}$ = 5 and the mean of the y series $\bar{y}$ =130, hence we can use the formula (2.1)

Calculation of correlation coefficient

x	**y**	**(x – $\bar{x}$)= x – 5**	**(y – $\bar{y}$) = y – 130**	**(x – $\bar{x}$)^2**	**(y – $\bar{y}$)^2**	**(x – $\bar{x}$) (y – $\bar{y}$)**
9	120	4	– 10	16	100	– 40
8	130	3	0	9	0	0
5	90	0	– 40	0	1600	0
1	150	– 4	20	16	400	– 80
4	160	– 1	30	1	900	– 30
2	100	– 3	– 30	9	900	90
3	140	– 2	10	4	100	– 20
6	80	1	– 50	1	2500	– 50
7	200	2	70	4	4900	140
				Σ(x – $\bar{x}$)^2= 60	**Σ(y – $\bar{y}$)^2= 11400**	**Σ(x – $\bar{x}$) (y – $\bar{y}$)= 10**

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} = \frac{10}{\left[60x11400\right]^{\frac{1}{2}}}$$

r(X,Y) = 0.012

Therefore, the correlation coefficient between the C.N.S. malformation rate and water hardness is 0.012.

Example 3: Find the product moment correlation for the following data

X	57	62	60	57	65	60	58	62	56
Y	71	70	66	70	69	67	69	63	70

Solution:

X	**Y**	**XY**	**X^2**	**Y^2**
57	71	4047	3249	5041
62	70	4340	3844	4900
60	66	3960	3600	4356
57	70	3990	3249	4900
65	69	4485	4225	4761
60	67	4020	3600	4489
58	69	4002	3364	4761
62	63	3906	3844	3969
56	70	3920	3136	4900
537	**615**	**36670**	**32111**	**42077**

Thus we have, n = 9, ∑X = 537, ∑Y = 615, ∑XY = 36670, $\sum X^2$ = 32111, $\sum Y^2$ = 42077

$$r(X,Y) = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - (\bar{x})(\bar{y})}{\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2}\ \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - (\bar{y})^2}} = -0.414$$

Example 4: A computer operator while calculating the coefficient of correlation between two variables X and Y for 25 pairs of observations obtained the following constants: ∑X = 125, ∑Y = 100, ∑XY = 508, $\sum X^2$ = 650, $\sum Y^2$ = 460. However it was later discovered at the time

of checking that he had copied two pairs as (6,14) and (8,6) while the correct pairs were (8,12) and (6,8). Obtain the correct correlation coefficient.

Solution:

The formula involved with the given data is, $$r(X,Y)=\frac{\frac{\sum_{i=1}^{n}x_iy_i}{n}-(\bar{x})(\bar{y})}{\sqrt{\frac{\sum_{i=1}^{n}x_i^{\,2}}{n}-(\bar{x})^{\,2}}\ \sqrt{\frac{\sum_{i=1}^{n}y_i^{\,2}}{n}-(\bar{y})^{\,2}}}$$

The Corrected $\sum X$ = Incorrect $\sum X$ – (6+8) + (8+6) = 125

Corrected $\sum Y$ = Incorrect $\sum Y$ – (14+6) + (12+8) = 100

Corrected $\sum X^2$ = Incorrect $\sum X^2$ – (6^2+8^2) + (8^2+6^2) = 650

Corrected $\sum Y^2$ = Incorrect $\sum Y^2$ – (14^2+6^2) + (12^2+8^2) = 436

Corrected $\sum XY$ = Incorrect $\sum XY$ – (84+48) + (96+48) = 520

Now the correct value of correlation coefficient is, $$r(X,Y)=\frac{\frac{520}{25}-(5\times 4)}{\sqrt{\frac{650}{25}-5^2}\sqrt{\frac{436}{25}-4^2}}$$ = 0.67

Partial Correlation Coefficient:

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variables, with the effect of the most of the variables eliminated.

Let $r_{12.3}$ be the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, then

$$r_{12.3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{\left(1-r_{13}^{\,2}\right)\left(1-r_{23}^{\,2}\right)}}$$

Similarly,

$$r_{13.2}=\frac{r_{13}-r_{12}r_{23}}{\sqrt{\left(1-r_{12}^{\,2}\right)\left(1-r_{23}^{\,2}\right)}}$$

where $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3 keeping X_2 constant.

$$r_{23.1}=\frac{r_{23}-r_{12}r_{13}}{\sqrt{\left(1-r_{12}^{\,2}\right)\left(1-r_{13}^{\,2}\right)}}$$

where $r_{23.1}$ is the coefficient of partial correlation between X_2 and X_3 keeping X_1 constant.

Problems:

1. If $r_{12} = 0.8$, $r_{13} = 0.4$ and $r_{23} = 0.56$, find the value of $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^{\,2}\right)\left(1 - r_{23}^{\,2}\right)}}$$

Substituting the given values,

$$r_{12.3} = \frac{0.8 - 0.4 \times 0.56}{\sqrt{1-(0.4)^2}\sqrt{1-(0.56)^2}}$$

$$= 0.7586$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1 - r_{12}^{\,2}\right)\left(1 - r_{23}^{\,2}\right)}}$$

$$= \frac{0.4 - (0.8)(0.56)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.56)^2}}$$

$$= -0.0966$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{\left(1 - r_{12}^{\,2}\right)\left(1 - r_{13}^{\,2}\right)}}$$

$$= \frac{0.56 - (0.8)(0.4)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.4)^2}}$$

$$= 0.4364$$

2. The correlation between a general intelligence test and school achievement in a group of children from 6 to 15 years old is 0.80. The correlation between the general intelligence test and age in the same group is 0.70 and the correlation between school achievement and age is 0.60. What is the correlation between general intelligence and school achievement in children of the same age?

Solution:

Let X_1 denote general intelligence test.

X_2 denote school achievement.

X_3 denote age.

We are given $r_{12} = 0.8$, $r_{13} = 0.7$ and $r_{23} = 0.6$

We have to find $r_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

$$= \frac{0.8 - (0.7)(0.6)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.6)^2}}$$

$$= 0.6651$$

Multiple Correlation

The coefficient of multiple linear correlation is represented by R, and it is common to add subscripts designating the variable involved. Thus $R_{1.23}$ would represent the coefficient of multiple linear correlation between X_1, on the one hand, and X_2 and X_3 on the other hand. The subscript of the dependent variable is always to the left of the point.

The coefficient of multiplication correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

The coefficient of multiple correlation lies between 0 and 1.

1. If $r_{12} = 0.09$, $r_{13} = 0.75$ and $r_{23} = 0.7$, find $R_{1.23}$

Solution:

We have to calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent, that is we have to find $R_{1.23}$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Substituting the given values,

$$R_{1.23} = \sqrt{\frac{0.9^2 + 0.75^2 - 2 \times 0.9 \times 0.75 \times 0.7}{1 - 0.7^2}}$$

$$= 0.9156$$

Spearman's Rank Correlation Coefficient

If X and Y are qualitative variables then Karl Pearson's coefficient of correlation will be meaningless. In this case, we use Spearman's rank correlation coefficient which is defined as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n\left(n^2 - 1\right)}$$ where d is the difference in ranks.

Problems:

1. The ranks of same 16 students in Mathematics and Physics are as follows. The numbers within brackets denote the ranks of the students in Mathematics and Physics. (1,1), (2,10), (3,3), (4,4), (5,5), (6,7), (7,2), (8,6), (9,8), (10,11) (11. 15), (12,9), (13,14), (14,12), (15,16), (16,13). Calculate the rank correlation coefficient for the proficiencies of this group in Mathematics and Physics.

Solution:

Ranks in Maths(X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics(Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
d = X – Y	0	–8	0	0	0	–1	5	2	1	–1	–4	3	–1	2	–1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Spearman's Rank Correlation Coefficient is given by, $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n\left(n^2 - 1\right)} = 1 - \frac{6 \times 136}{16\left(16^2 - 1\right)}$ = 0.8

2. The coefficient of rank correlation between the marks in Statistics and Mathematics obtained by a certain group of students is 2/3 and the sum of the squares of the differences in ranks is 55. Find the number of students in the group.

Solution: Spearman's rank correlation coefficient is given by $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n\left(n^2 - 1\right)}$

Here ρ = 2/3, $\sum d^2 = 55$, N = ? Therefore $\frac{2}{3} = 1 - \frac{6 \times 55}{n(n^2 - 1)}$ Solving the above equation we get n = 10.

Repeated Ranks:

If any two or more individuals are equal in the series then Spearman's formula for calculating the rank correlation coefficients breaks down. In this case, common ranks are given to the repeated ranks. This common rank is the average of the ranks which these items would have assumed if they are slightly different from each other and the next item will get the rank next the ranks already assumed. As a result of this, following adjustment is made in the formula: add the factor $\frac{m(m^2 - 1)}{12}$ to $\sum d^2$ where m is the number of items an item is repeated. This correction factor is to be added for each repeated value.

3. Obtain the rank correlation coefficient for the following data:

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Solution:

X	**Y**	**Rank X**	**Rank Y**	**D = X – Y**	**D^2**
68	62	4	5	– 1	1
64	58	6	7	– 1	1
75	68	2.5	3.5	– 1	1
50	45	9	10	– 1	1
64	81	6	1	5	25
80	60	1	6	– 5	25
75	68	2.5	3.5	– 1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					72

In X series 75 is repeated twice which are in the positions 2nd and 3rd ranks. Therefore common ranks 2.5 (which is the average of 2 and 3) is given for each 75. The corresponding correction factor is $C.F = \frac{2(2^2 - 1)}{12} = \frac{1}{2}$. Also in the X series 64 is repeated

thrice which are in the position 5th, 6th and 7th ranks. Therefore, common ranks 6(which is the average of 5, 6 and 7) is given for each 64. The corresponding correction factor is $C.F = \frac{3(3^2-1)}{12} = 2$. Similarly, in the Y series, 68 is repeated twice which are in the positions 3rd and 4th ranks. Therefore, common ranks(which is the average of 3 and 4) is given for each 68. The corresponding correction factor is $C.F = \frac{2(2^2-1)}{12} = \frac{1}{2}$. Rank correlation coefficient is $\rho = 1 - \frac{6\left(\Sigma d^2 + TotalCorrectionFactor\right)}{n\left(n^2-1\right)} = 1 - \frac{6\left(72 + \frac{1}{2} + 2 + \frac{1}{2}\right)}{10(10-1)}$ = 0.5454.

Regression Analysis

Regression analysis helps us to estimate or predict the value of one variable from the given value of another. The known variable(or variables) is called independent variable(s). The variable we are trying to predict is the dependent variable.

Regression equations

Prediction or estimation of most likely values of one variable for specified values of the other is done by using suitable equations involving the two variables. Such equations are known as Regression Equations

Regression equation of y on x:

$y - \bar{y} = b_{yx}(x - \bar{x})$ where y is the dependent variable and x is the independent variable and b_{yx} is given by

$$b_{yx} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2} \quad \text{or} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\frac{\sum_{i=1}^{n}xy}{n} - (\bar{x}\bar{y})}{\frac{\sum_{i=1}^{n}x^2}{n} - (\bar{x})^2}$$

Regression equation of x on y:

$x - \bar{x} = b_{xy}(y - \bar{y})$ where y is the dependent variable and x is the independent variable and b_{yx} is given by

$$b_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(y-\bar{y})^2} \quad \text{or} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\frac{\sum_{i=1}^{n}xy}{n} - (\bar{x}\bar{y})}{\frac{\sum_{i=1}^{n}y^2}{n} - (\bar{y})^2}$$

b_{yx} and b_{xy} are called as regression coefficients of y on x and x on y respectively.

Relation between correlation and regression coefficients:

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ $b_{yx} \cdot b_{xy} = r\frac{\sigma_y}{\sigma_x} \; r\frac{\sigma_x}{\sigma_y} = r^2$ Hence

$r = \pm\sqrt{b_{yx} b_{xy}}$

Note: In the above expression the components inside the square root is valid only when b_{yx} and b_{xy} have the same sign. Therefore the regression coefficients will have the same sign.

Problems:

1. In trying to evaluate the effectiveness of its advertising campaign a company compiled the following information. Calculate the regression line of sales on advertising.

Year	1980	1981	1982	1983	1984	1985	1986	1987
Advertisement in 1000 rupees	12	15	15	23	24	38	42	48
Sales in lakhs of rupees	5	5.6	5.8	7.0	7.2	88	9.2	9.5

Solution : Let x be advertising amount and y be the sales amount.

Here, n = 8, $\bar{x} = \frac{217}{8} = 27.1$, $\bar{y} = \frac{58.1}{8} = 7.26$

We know that, Regression equation of y on x is given by $y - \bar{y} = b_{yx}(x - \bar{x})$ where

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\frac{\sum_{i=1}^{n} xy}{n} - (\bar{x}\bar{y})}{\frac{\sum_{i=1}^{n} x^2}{n} - (\bar{x})^2}$$

X	**Y**	**X^2**	**XY**
12	5	144	60
15	5.6	225	84
15	5.8	225	87
23	7.0	529	161
24	7.2	576	172.8
38	8.8	1444	334.4
42	9.2	1764	386.4
48	9.5	2304	456
217	**58.1**	**7211**	**1741.6**

Therefore $b_{yx} = 0.125$ Substituting this value in the y on x equation, we get,

y – 7.26 = 0.125(x – 27.1) Therefore the required equation of Sales on Advertisement is y = 3.87 + 0.125 x

2. In a study of the effect of a dietary component on plasma lipid composition, the following ratios were obtained on a sample of experimental anumals

Measure of dietary component (X)	1	5	3	2	1	1	7	3
Measure of plasma lipid level (Y)	6	1	0	0	1	2	1	5

(i) obtain the two regression lines and hence predict the ratio of plasma lipid level with 4 dietary component.

(ii) find the correlation coefficient between X and Y

Solution: (i)

X	**Y**	**XY**	**X^2**	**Y^2**
1	6	6	1	36
5	1	5	25	1
3	0	0	9	0
2	0	0	4	0
1	1	1	1	1
1	2	2	1	4
7	1	7	49	1
3	5	15	9	25
23	**16**	**36**	**99**	**68**

Here n = 8 $\bar{x}$ = 2.875 $\bar{y}$ = 2 The Regression equation of y on x is given by y – $\bar{y}$ = b_{yx} (x – $\bar{x}$)

Where $b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\frac{\sum_{i=1}^{n} xy}{n} - (\bar{x}\bar{y})}{\frac{\sum_{i=1}^{n} x^2}{n} - (\bar{x})^2}$ $b_{yx} = -0.304$

Hence the regression equation of y on x is

$y - 2 = -0.304(x - 2.875)$

(i.e) $y = 2.874 - 0.304\,x$

when x = 4 (measure of dietary component) the plasmid lipid level is

$y = 2.874 - 0.304\,(4)$

$y = 1.658$

The Regression equation of x on y is given by $x - \bar{x} = b_{xy}(y - \bar{y})$

Where $b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\frac{\sum_{i=1}^{n} xy}{n} - (\bar{x}\bar{y})}{\frac{\sum_{i=1}^{n} y^2}{n} - (\bar{y})^2}$ $\qquad b_{xy} = -0.278$

Hence the regression equation of x on y is

$x - 2.875 = -0.278(y - 2)$

(i.e) $x = 3.431 - 0.278\,y$

(ii) The correlation coefficient between x and y is given by

$$r = \pm\sqrt{b_{yx}b_{xy}}$$

$$r = \pm\sqrt{-0.304 \times -0.278} = \pm 0.291$$

3. From the data given below find (i) two regression lines (ii) coefficient of correlation between marks in Physics and marks in Chemistry (iii) most likely marks in Chemistry when marks in Physics is 78 (iv) most likely marks in Physics when marks in Chemistry is 92

Marks in Physics (X)	72	85	91	85	91	89	84	87	75	77
Marks in Chemistry (Y)	76	92	93	91	93	95	88	91	80	81

Solution:

(i)

X	**Y**	**X^2**	**Y^2**	**XY**
72	76	5184	5776	5472
85	92	7225	8464	7820
91	93	8281	8649	8463

85	91	7225	8281	7735
91	93	8281	8649	8463
89	95	7921	9025	8455
84	88	7056	7744	7395
87	91	7569	8281	7917
75	80	5625	6400	6000
77	81	5929	6561	6237
836	**880**	**70296**	**77830**	**73957**

Here n = 10 $\bar{x}$ = 83.6 $\bar{y}$ = 88

The Regression equation of y on x is given by $y - \bar{y} = b_{yx}(x - \bar{x})$

Where

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\frac{\sum_{i=1}^{n} xy}{n} - (\bar{x}\bar{y})}{\frac{\sum_{i=1}^{n} x^2}{n} - (\bar{x})^2}$$

$$b_{yx} = 0.949$$

Hence the regression equation of y on x is

$$y - 88 = 0.949(x - 83.6)$$

(i.e) $y = 8.6 + 0.949x$

The Regression equation of x on y is given by $x - \bar{x} = b_{xy}(y - \bar{y})$

Where

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\frac{\sum_{i=1}^{n} xy}{n} - (\bar{x}\bar{y})}{\frac{\sum_{i=1}^{n} y^2}{n} - (\bar{y})^2}$$

$$b_{xy} = 0.990$$

Hence the regression equation of x on y is

$$x - 83.6 = 0.990(y - 88)$$

(i.e) $x = -\ 3.5 + 0.990\ y$

(ii) The correlation coefficient between x and y is given by

$$r = \pm\sqrt{b_{yx} b_{xy}}$$

$$r = \pm\sqrt{0.949 \times 0.990} = \pm 0.969$$

(iii) To find the most likely marks in Chemistry when marks in Physics is 78, we have to use the regression equation of y on x given by

$$y = 8.6 + 0.949\ x$$

Substituting the value of x as 78 in the above equation, we get,

$$y = 8.6 + 0.949\ (78)$$

$$y = 73.85$$

Hence the marks in Chemistry is 82.62

(iv) To find the most likely marks in Physics when marks in Chemistry is 92, we have to use the regression equation of x on y given by

$$x = -\ 3.5 + 0.990\ y$$

Substituting the value of y as 92 in the above equation, we get,

$$x = -\ 3.5 + 0.990\ (92)$$

$$x = 87.58$$

Hence the marks in Physics is 87.58

4. For a given series of values, the following data were obtained, $\bar{x}$ =36, $\bar{y}$ =85, σ_x = 11, σ_y = 8 and r = 0.66. Find (i) two regression equations (ii) estimation of x when y = 75.

Solution:

We have $b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.66 \times \frac{8}{11} = 0.4799$

and $b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.66 \times \frac{11}{8} = 0.9075$

(i) The Regression equation of y on x is given by

$y - \bar{y} = b_{yx}(x - \bar{x})$

$y - 85 = 0.4799\,(x - 36)$

(i.e.) $y = -17.28 + 0.4799\,x$

The Regression equation of x on y is given by

$x - \bar{x} = b_{xy}(y - \bar{y})$

$x - 36 = 0.9075(y - 85)$

(i.e.) $x = -41.35 + 0.9075\,y$

(ii) To estimate the value of x when y = 75, we use the regression line of x on y

$x = -41.35 + 0.9075\,y$

Substituting y = 75, $x = -41.35 + 0.9075\,(75)$

Therefore $x = 29.9$

5. For a certain X and Y series which are correlated, the regression lines are 8x – 10y = – 66 and 40x – 18y = 214. Find (i) the correlation coefficient between them and (ii) the mean of the two series.

Solution:

The given regression equations are

$8x - 10y = -66$ ……(1)

$40x - 18y = 214$ ……(2)

(i) Let us suppose that the equation (1) is the equation of line of regression of y on x and (2) as the equation of the line of regression of x on y, after rewriting (1) and (2), we get

$y = \frac{66}{10} + \frac{8}{10}x$ which gives the value of $b_{yx} = \frac{8}{10}$

x = $\frac{214}{40} + \frac{18}{40}$ y which gives the value of $b_{xy} = \frac{18}{40}$

Now $r = \pm\sqrt{b_{yx} b_{xy}} = \pm\sqrt{\frac{8}{10} \times \frac{18}{40}} = \pm 0.6$

(ii) Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point ($\bar{x}$, $\bar{y}$) must satisfy the given two regression lines.

Therefore, $8\ \bar{x} - 10\ \bar{y} = -66$

$40\ \bar{x} - 18\ \bar{y} = 214$

Solving the above two equations we get $\bar{x} = 13$ and $\bar{y} = 17$

Important Note: In the above problem in part (i), if we take equation (1) as the line of regression of x on y, we get, x = $-\frac{66}{8} + \frac{10}{18}$ y , and hence $b_{xy} = \frac{10}{8}$

and if we take equation (2) as the line of regression of y on x, we get,

y = $-\frac{214}{18} + \frac{40}{18}$ x and hence $b_{yx} = \frac{40}{18}$

Therefore, $r = \pm\sqrt{b_{yx} b_{xy}} = \pm\sqrt{\frac{10}{8} \times \frac{40}{18}} = \pm 1.67$

But the value of r cannot exceed unity. Hence the assumptions that line (1) is line of regression of x on y and the line (2) is line of regression of y on x are wrong.

Fitting curves by Method of Least Squares

Curve Fitting: Let (x_i, y_i); i = 1, 2, …, n be a given set of n pairs of values, X being independent variable and Y being the dependent variable. The general problem in curve fitting is to find, if possible, an analytic expression of the form y = f(x), for the functional relationship suggested by the given data.

Fitting a straight line

Let y = a + bx be the equation of the line to be fitted. To estimate the values of a and b we have, the following normal equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2}$$

Problems:

1. Fit a straight line to the following data:

X	1	2	3	4	6	8
Y	2.4	3	3.6	4	5	6

Solution : Let the straight line to be fitted is y = a + bx

X	**Y**	**XY**	**X^2**
1	2.4	2.0	1
2	3	6.0	4
3	3.6	10.8	9
4	4	16.0	16
6	5	30.0	36
8	6	48.0	64
24	**24**	**113.2**	**130**

Using the normal equations, $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$

$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2}$ we get,

24 = 6a + 24b and

113.2 = 24a + 130b

Solving above two equations, we get

a = 1.976 and b = 0.506

2. Fit a straight line of the form y = a + bx for the following data and estimate the value of y when x is 40

X	2	4	6	10	20	24
Y	6	8	13	12	35	42

Solution: Here n = 6

X	**Y**	**XY**	**X^2**
2	6	12	4
4	8	24	16
6	13	78	36
10	12	120	100
20	35	700	400
24	42	1008	576
66	**116**	**1950**	**1132**

Using the normal equations, $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2}$$ we get,

116 = 6 a + 66 b and

1950 = 66 a + 1132

Solving the above two equations, we get

a = 1.073 and

b = 1.66

Now to estimate the value of y when x is 40, we substitute the value of x in the fitted equation

y = a + b x

(i.e.) y = 1.07 + 1.66 x

= 1.07 + 1.66 x 40

y = 67.47

Fitting a parabola

Let $y = a + bx + cx^2$ be the equation of the line to be fitted. To estimate the values of a and b and c, we have, the following normal equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\,2}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3}$$

$$\sum_{i=1}^{n} x_i^{\,2} y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i^{\,3} + c\sum_{i=1}^{n} x_i^{\,4}$$

Problems:

1. Fit a parabola to the following data:

X	0	1	2	3	4
Y	1	1.8	1.3	2.5	6.3

Solution: Let $y = a + bx + cx^2$ be the second degree parabola to be fitted, n = 5

X	**Y**	**X^2**	**X^3**	**X^4**	**XY**	**X^2Y**
0	1.0	0	0	0	0	0
1	1.8	1	1	1	1.8	1.8
2	1.3	4	8	16	2.6	5.2
3	2.5	9	27	81	7.5	22.5
4	6.3	16	64	256	25.2	100.8
10	**12.9**	**30**	**100**	**354**	**37.1**	**130.3**

Using normal equations $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\,2}$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2 + c\sum_{i=1}^{n} {x_i}^3$$

$$\sum_{i=1}^{n} {x_i}^2 y_i = a\sum_{i=1}^{n} {x_i}^2 + b\sum_{i=1}^{n} {x_i}^3 + c\sum_{i=1}^{n} {x_i}^4$$ we get,

12.9 = 5a + 10b + 30c

37.1 = 10a + 30b + 100c

130.3 = 30a + 100b + 354c

Solving the above equations, we get

a = 1.42

b = – 1.07

c = 0.55.

Thus the required equation of parabola is $y = 1.42 – 1.07x + 0.55x^2$

2. Fit a parabola to the following data and estimate y when x is 6

X	1	3	4	5	7
Y	2	3	6	15	39

Solution: Let $y = a + bx + cx^2$ be the second degree parabola to be fitted, n = 5

X	**Y**	**X^2**	**X^3**	**X^4**	**XY**	**X^2Y**
1	2	1	1	1	2	2
3	3	9	27	81	9	27
4	6	16	64	256	24	96
5	15	25	125	625	75	375
7	39	49	343	2401	273	1911
20	**65**	**100**	**560**	**3364**	**383**	**2411**

Using normal equations $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} {x_i}^2$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3}$$

$$\sum_{i=1}^{n} x_i^{\,2} y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i^{\,3} + c\sum_{i=1}^{n} x_i^{\,4}$$ we get,

65 = 5a + 20b + 100c

383 = 20a + 100b + 560c

2411 = 100a + 560b + 3364c

Solving the above equations, we get

a = 6.54

b = – 5.93

c = 1.51.

Thus the required equation of parabola is $y = 6.54 – 5.93x + 1.51x^2$

Now to estimate the value of y when x is 6, we substitute the value of x in the fitted equation

$y = a + bx + cx^2$

$= 6.54 – 5.93 \times 6 + 1.51 \times 6^2$

y = 25.32

EXERCISE PROBLEMS:

1. State Spearman's formula to find rank correlation coefficient for repeated ranks.

2. State any two properties of correlation coefficient.

3. For the given bivariate data, (a) Fit a regression line y on x and predict the value of y when x = 5. (b) Fit a regression line x on y and predict the value of x when y = 2.5 (c) Calculate Karl pearson's coefficient of correlation.

x	1	5	3	2	1	1	7	3
y	6	1	0	0	1	2	1	5

4. Ten competitors in a beauty contest are ranked by 3 judges in the following order

I Judge	1	6	5	10	3	2	4	9	7	8
II Judge	3	5	8	4	7	10	2	1	6	9
III Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to discuss which pair of judges have the Nearest approach to common taste of beauty.

5. Find the rank correlation coefficient for the data given below:-

x	70	75	80	60	70	71	82	83	85
y	60	62	61	70	61	72	75	61	59

6. Write down the distinguishing features of correlation and regression.

7. Find the most likely price in Bombay corresponding to the price of Rs. 70 at Calcutta from the following:

	Calcutta	Bombay
Average price	65	67
Standard deviation	2.5	3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8.

8. Obtain the rank correlation coefficient for the following data:

X :	68	64	75	50	64	80	75	40	55	64
Y :	62	58	68	45	81	60	68	48	50	70

9. Fit a parabola of second degree to the following data:

X :	1	2	3	4	6	8
Y :	2.4	3	3.6	4	5	6

10. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of X = 9. Regression equations: 8x – 10y + 66 = 0, 40x – 18y = 214. What were

(i) the mean values of X and Y .

(ii) the correlation coefficient between X and Y and

(iii) the standard deviation of Y?

11. Calculate the correlation coefficient for the following heights (in inches) of fathers(X) and their sons(Y):

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71