1. PROBABILITY AND EXPECTATION

STRUCTURE

1.1. INTRODUCTION

The word 'Probability' and 'Chance' are quite familiar to everyone. Many a times, we come across statements like, "Probably it may rain today", "chances of hitting the target are very few". "It is possible that he may top the examination". In the above statements, the probably, chances, possible, etc. convey the sense of uncertainty about the occurrence of some event. Ordinarily, it appears that there cannot be any exact measurement for these uncertainties, but in Mathematical Statistics, we have methods for calculating the degree of certainty of events in numerical value, under certain conditions. When, we perform experiments in science and engineering, repeatedly under identical conditions, we get almost the same result. There also exist experiment in which the outcome may be different even if the experiment is performed under identical conditions. In such experiments, the outcome of each experiment depends on chance.

1.2. SOME IMPORTANT DEFINITIONS

Experiment : Any operation that results in two or more outcomes is called an experiment, and performing of an experiment is called trial.

NOTES

Random Experiment : A random experiment is defined as an experiment in which all possible outcomes are known and which can be repeated under identical conditions but it is not possible to predict the outcome of any particular trial in advance.

e.g. Tossing a coin or throwing a die is random experiment.

Sample Space : The sample space of a random experiment is defined as the set of all possible outcomes of the experiment. The possible outcomes are called sample points. The sample space is generally denoted by the letter S.

e.g. In throwing a fair die, sample space is S = {1, 2, 3, 4, 5, 6}. In tossing of two unbiased coins sample space is S = {HH, HT, TH, TT}.

Event : Any subset of the sample space is defined as an event. An event is called an elementary (or simple) event if it contains only one sample point. In the experiment of throwing a die, the event A of getting 2 is a simple event. We write A = {2}. Also an event is called an impossible event if it can never occur. In the above experiment, event B = {7} of getting 7 is an impossible event. An event which is sure to occur is called a certain event.

e.g. In throwing a die, the event of getting a number less than 7 is a certain event.

Exhaustive Events : The total number of all possible outcomes in any trial are known as exhaustive events or cases.

e.g. In tossing a coin, there are two exhaustive events, head and tail. In throwing a die, there are 6 exhaustive cases, any one of the six faces may turn up.

Note. In throwing n dice, the exhaustive cases are 6^n.

Equally Likely Events : Events are said to be equally likely, if there is no reason to expect any one in preference to any other.

e.g. If we draw a card from a well-shuffled pack, we may get any card, then the 52 different cases are equally likely.

Favourable Events : The events which ensure the required happening, are said to be favourable events.

e.g. In throwing a die, the number of cases favourable to the appearance of a multiple of 2 are three *viz.* 2, 4 and 6. In drawing two cards from a pack of 52 cards, the number of cases favourable to drawing 2 aces is ${}^4C_2$.

Independent Events : Events are said to be independent if the happening (or non-happening) of one event is not affected by the happening (or non-happening) of others.

e.g. In case a card is drawn from a pack of well shuffled cards and is not replaced, then the second draw of the card is dependent on the first draw. However, if the first card drawn is replaced before drawing the second card, the result of the second draw is independent of the first draw.

Mutually Exclusive Events : Two events are said to be mutually exclusive if they cannot occur together *i.e.,* if one occurs then other cannot.

e.g. In tossing a coin, the events head and tail are mutually exclusive, since if the outcome is tail, the possibility of getting head in the same trial is ruled out.

Compound Events : Events obtained by combining together two or more elementary events are known as the compound events.

e.g. In throwing a die, getting 5 or 6 is called a compound event.

Mathematical (or Classical) Definition of Probability : If an event can happen in n ways which are equally likely, exhaustive and mutually exclusive and out

of these n ways, m ways are favourable to an event A, then the probability of happening of A is given by

$$p \text{ or } P(A) = \frac{m}{n}$$

If A happens in m ways, it will fail in $(n - m)$ ways so that the probability of its failure

$$q \quad \text{or} \quad P(\bar{A}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p$$

$$\Rightarrow \qquad p + q = 1 \quad i.e., \quad P(A) + P(\bar{A}) = 1$$

$$0 \le p \le 1 \; ; \; 0 \le q \le 1$$

If P(A) = 1, then A is called a certain event. If P(A) = 0, then A is called an impossible event.

Statistical (or Empirical) Definition of Probability : If in n trials, an event A happens m times then the probability of happening A is given by

$$p \text{ or } P(A) = \lim_{n \to \infty} \frac{m}{n}.$$

THEOREMS ON PROBABILITY

1.3. ADDITION THEOREMS ON PROBABILITY

1.3.1. Theorem 1 *(Addition Theorem for Two Events)*

If A and B are two events associated with a random experiment, then

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

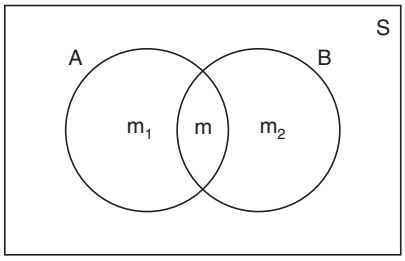
Proof. Let S be the sample space associated with the given random experiment. Suppose the experiment results in n mutually exclusive ways. Then S contains n elementary events.

Let m_1, m_2 and m be the number of elementary events favourable to A, B and $A \cap B$ respectively. Then,

$$P(A) = \frac{m_1}{n}, \; P(B) = \frac{m_2}{n}$$

and

$$P(A \cap B) = \frac{m}{n}.$$



The number of elementary events favourable to A only is $m_1 - m$. Similarly, the number of elementary events favourable to B only is $m_2 - m$. Since m elementary events are favourable to both A and B, therefore, the number of elementary events favourable to A or B or both *i.e.*, $A \cup B$ is

$$m_1 - m + m_2 - m + m = m_1 + m_2 - m.$$

So,

$$P(A \cup B) = \frac{m_1 + m_2 - m}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{m}{n}$$

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

NOTES

Corollary : If A and B are mutually exclusive events, then $P(A \cap B) = 0$, therefore,

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

This is the addition theorem for mutually exclusive events.

1.3.2. Theorem 2 *(Addition Theorem for three events)*

If A, B and C are three events associated with a random experiment, then

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

Proof. Let $D = B \cup C$, then

$$P(A \cup B \cup C) = P(A \cup D) = P(A) + P(D) - P(A \cap D) \quad ...(1) \text{ (by Th. 1)}$$

Now, $A \cap D = A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$\therefore$ $P(A \cap D) = P[(A \cap B) \cup (A \cap C)]$

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

$$= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \quad ...(2) \text{ (by Th. 1)}$$

$[\because \quad (A \cap B) \cap (A \cap C) = A \cap B \cap C]$

Also $P(D) = P(B \cup C) = P(B) + P(C) - P(B \cap C)$ ...(3)

From (1), (2) and (3), we get

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

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

Corollary : If A, B and C are mutually exclusive events, then

$$P(A \cap B) = P(B \cap C) = P(A \cap C) = P(A \cap B \cap C) = 0$$

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

This is addition theorem for three mutually exclusive events.

1.4. CONDITIONAL PROBABILITY

Let A and B be two events associated with a random experiment. Then, the probability of occurrence of A under the condition that B has already occurred and $P(B) \neq 0$, is called the conditional probability and is denoted by P(A/B).

Thus, P(A/B) = Probability of occurrence of A given that B has already occurred.

Similarly, P(B/A) = Probability of occurrence of B given that A has already occurred.

1.5. MULTIPLICATION THEOREMS ON PROBABILITY

1.5.1. Theorem 1

If A and B are two events associated with a random experiment, then

$$P(A \cap B) = P(A)\, P(B/A), \quad \text{if} \quad P(A) \neq 0$$

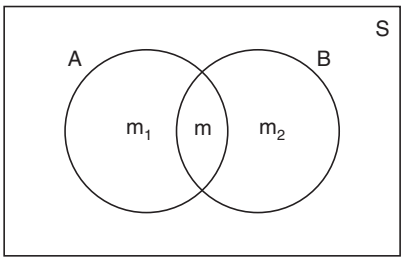
or

$$P(A \cap B) = P(B)\, P(A/B), \quad \text{if} \quad P(B) \neq 0$$

Proof. Let S be the sample space associated with the given random experiment. Suppose S contains n elementary events. Let m_1, m_2 and m be the number of elementary events favourable to A, B and A ∩ B respectively. Then

$$P(A) = \frac{m_1}{n},\ P(B) = \frac{m_2}{n} \quad \text{and} \quad P(A \cap B) = \frac{m}{n}.$$

Since m_1 elementary events are favourable to A out of which m are favourable to B, therefore,



$$P(B/A) = \frac{m}{m_1}.$$

Similarly, $P(A/B) = \frac{m}{m_2}$

Now, $P(A \cap B) = \frac{m}{n} = \frac{m}{m_1} \cdot \frac{m_1}{n}$

$= P(B/A)\ .\ P(A)$ ...(1)

and $P(A \cap B) = \frac{m}{n} = \frac{m}{m_2} \cdot \frac{m_2}{n} = P(A/B)\ P(B)$ ...(2)

Note 1. From (1) and (2) in the above theorem, we find that

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \quad \text{and} \quad P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$P(A \cap B)$ is also written as $P(AB)$.

2. For three events A, B, C

$P(A \cap B \cap C) = P(ABC)$

= Probability of the simultaneous occurrence of events A, B and C

$= P(A)\ P(B/A)\ P(C/AB)$

$= P(A)\ P(B/A)\ P(C/(A \cap B))$

If $A_1, A_2, ..., A_n$ are n events, then

$P(A_1 \cap A_2 \cap \ ... \cap A_n) = P(A_1)\ P(A_2/A_1)\ P(A_3/A_1 \cap A_2)\ ...\ P(A_n/A_1 \cap A_2 \cap ... \cap A_{n-1})$

1.5.2. Multiplication Theorems For Independent Events

Theorem 1. *If A and B are independent events associated with a random experiment, then*

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

Proof. By multiplication theorem, we have

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

Since A and B are independent events, therefore, $P(B/A) = P(B)$

Hence, $P(A \cap B) = P(A)\ P(B)$

Theorem 2. *If $A_1, A_2, ..., A_n$ are independent events associated with a random experiment, then*

$$P(A_1 \cap A_2 \cap \ A_3 \ ... \cap A_n) = P(A_1)\ P(A_2)\ ...\ P(A_n)$$

Proof. By multiplication theorem, we have

$P(A_1 \cap A_2 \cap A_3 \cap \ ... \cap A_n)$

$= P(A_1)\ P(A_2/A_1)\ P(A_3/A_1 \cap A_2)\ ...\ P(A_n/A_1 \cap A_2 \cap ... \cap A_{n-1})$

Since $A_1, A_2, \ ..., A_{n-1}, A_n$ are independent events, therefore,

$P(A_2/A_1) = P(A_2),\ P(A_3/A_1 \cap A_2) = P(A_3),\ ...,\ P(A_n/A_1 \cap A_2 \cap ... \cap A_{n-1})$

$= P(A_n)$

Hence, $P(A_1 \cap A_2 \cap A_3 \cap ... \cap A_n) = P(A_1)\ P(A_2)\ ...\ P(A_n)$

1.6. ADDITION THEOREM FOR INDEPENDENT EVENTS

1.6.1. Theorem

NOTES

If $A_1, A_2, ..., A_n$ are n independent events associated with a random experiment, then

$$P(A_1 \cup A_2 \cup ... \cup A_n) = 1 - P(\overline{A}_1)\, P(\overline{A}_2) ... P(\overline{A}_n).$$

Proof. We have $P(A_1 \cup A_2 \cup ... \cup A_n) = 1 - P(\overline{A_1 \cup A_2 \cup ... \cup A_n})$

$$= 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap ... \cap \overline{A}_n)$$

$$= 1 - P(\overline{A}_1)\, P(\overline{A}_2) ... P(\overline{A}_n)$$

($\because$ $A_1, A_2, ..., A_n$ are independent events, therefore, so are $\overline{A}_1, \overline{A}_2, ..., \overline{A}_n$)

SOLVED EXAMPLES

Example 1. *Find the probability of getting a tail in throw a coin.*

Solution. Clearly the sample space S = {H, T}

Event of getting tail E = {T}

Clearly $n(E) = 1$ and $n(S) = 2$

$\therefore$ Probability of getting a tail is given by

$$P(E) = \frac{n(E)}{n(S)} = \frac{1}{2}$$

or If E is the required event, then E = {T}

Hence, $$P(E) = \frac{\text{No. of cases favourable to E}}{\text{Total number of cases}} = \frac{1}{2}.$$

Example 2. *Three coins are tossed, find the probability of getting at least two heads.*

Solution. Clearly the sample space

S = {HHH, HHT, HTH, THH, THT, TTH, HTT, TTT}

If E is the required event, then

E = {HHH, HHT, HTH, THH}

$$P(E) = \frac{\text{No. of cases favourable to E}}{\text{Total number of cases}}$$

$$= \frac{4}{8} = \frac{1}{2}.$$

Example 3. *If there are two children in a family, find the probability that there is at least one girl in the family.*

Solution. Let S be the sample space, then

S = {BB, BG, GB, GG},

where B and G stand for 'Boy' and 'Girl' respectively.

If E is the required event, then

A = {BG, GB, GG}

$$P(E) = \frac{3}{4}.$$

Example 4. *What is the chance that a leap-year, selected at random, will contain 53 Fridays ?*

Solution. There are 366 days in a leap-year and we can write $366 = (7 \times 52) + 2$. This means that the leap year will contain at least 52 Fridays. The possible combinations for the remaining two days can be made as follows :

(*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.

Of these seven likely cases only (*v*) and (*vi*) are favourable.

Hence, the required probability $= \frac{2}{7}$.

Example 5. *What is the probability of getting an even number in the throw of an unbiased die ?*

Solution. Clearly, there are 6 equally likely possible outcomes 1, 2, 3, 4, 5, 6. Hence, the sample space $S = \{1, 2, 3, 4, 5, 6\}$

Let E be the required event, then we have

$$E = \{2, 4, 6\}$$

Hence, $$P(E) = \frac{3}{6} = \frac{1}{2}.$$

Example 6. *A bag contains 7 red, 12 white and 4 green balls. What is the probability that*

(*i*) *3 balls drawn are all white and*

(*ii*) *3 balls drawn are one of each colour.*

Solution. Total balls are = 7 + 12 + 4 = 23

3 balls out of these 23 balls can be drawn in

$$^{23}C_3 = \frac{23 \times 22 \times 21}{3 \times 2 \times 1} = 1771 \text{ ways}$$

$\therefore$ The sample space for this experiment contains 1771 sample point, *i.e.,* $n(S) = 1771$.

(*i*) Let E_1 = event that the 3 balls drawn are all white. Now 3 white balls can be drawn from 12 white balls in

$$^{12}C_3 = \frac{12 \times 11 \times 10}{3 \times 2 \times 1} = 220 \text{ ways}$$

$\therefore$ $$n(E_1) = 220$$

$\therefore$ $$P(E_1) = \frac{n(E_1)}{n(S)} = \frac{220}{1771}.$$

(*ii*) Let E_2 = event that three balls are one of each colour.

Now 1 red ball can be drawn out of 7 red balls in $^7C_1 = 7$ ways,

1 white ball can be drawn out of the 12 white balls in $^{12}C_1 = 12$ ways and 1 green ball can be drawn out of the 4 green balls in $^4C_1 = 4$ ways.

3 balls one of each colour can be drawn in $7 \times 12 \times 4 = 336$

$\therefore$ $$n(E_2) = 336$$

$\therefore$ $$P(E_2) = \frac{n(E_2)}{n(S)} = \frac{336}{1771}.$$

NOTES

Example 7. *From a pack of 52 cards three are drawn at random. Find the chance that they are a king, a queen and a knave.*

Solution. From a pack of 52 cards three can be drawn in ${}^{52}C_3$ ways. Thus, $n = {}^{52}C_3$.

There are 4 kings, 4 queens and 4 knaves. A king can be drawn in ${}^4C_1$ ways, a queen in ${}^4C_1$ ways and a knave in ${}^4C_1$ ways. Since each of these may be with drawn in ${}^4C_1 \times {}^4C_1 \times {}^4C_1$ ways.

$\therefore \quad m = {}^4C_1 \times {}^4C_1 \times {}^4C_1$

$\therefore$ Required probability $= \dfrac{{}^4C_1 \times {}^4C_1 \times {}^4C_1}{{}^{52}C_3} = \dfrac{4 \times 4 \times 4 \times 3 \times 2 \times 1}{52 \times 51 \times 50} = \dfrac{16}{5525}$.

Example 8. *A and B are two mutually exclusive events of an experiment. If P('not A') = 0.65, P(A ∪ B) = 0.65 and P(B) = p, find the value of p.*

Solution. By addition theorem for mutually exclusive events, we have

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

$$P(A \cup B) = 1 - P(\text{'not A'}) + P(B) \qquad [\because \quad P(A) = 1 - P(\overline{A})]$$

$$0.65 = 1 - 0.65 + p$$

$\Rightarrow \quad p = 0.30.$

Example 9. *The probability that at least one of the events A and B occurs is 0.6. If A and B occur simultaneously with probability 0.2, then find $P(\overline{A}) + P(\overline{B})$.*

Solution. We have $P(A \cup B) = 0.6$ and $P(A \cap B) = 0.2$

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

$$0.6 = P(A) + P(B) - 0.2$$

$$0.6 = 1 - P(\overline{A}) + 1 - P(\overline{B}) - 0.2 = 1.8 - [P(\overline{A}) + P(\overline{B})]$$

$\Rightarrow \quad P(\overline{A}) + P(\overline{B}) = 1.8 - 0.6 = 1.2.$

Example 10. *A, B, C are three mutually exclusive and exhaustive events associated with a random experiment. Find P(A), it being given that $P(B) = \dfrac{3}{2} P(A)$ and $P(C) = \dfrac{1}{2} P(B)$.*

Solution. Let $P(A) = p$. Then

$$P(B) = \frac{3}{2} P(A) \quad \Rightarrow \quad P(B) = \frac{3}{2} p$$

and

$$P(C) = \frac{1}{2} P(B) \quad \Rightarrow \quad P(C) = \frac{3}{4} p$$

Since A, B, C are mutually exclusive and exhaustive events associated with a random experiment, therefore,

$$A \cup B \cup C = S$$

$$P(A \cup B \cup C) = P(S) \quad \Rightarrow \quad P(A \cup B \cup C) = 1 \qquad [\because \quad P(S) = 1]$$

$$P(A) + P(B) + P(C) = 1$$

$$p + \frac{3}{2} p + \frac{3}{4} p = 1 \quad \Rightarrow \quad p = \frac{4}{13}.$$

Example 11. *A card is drawn from a pack of 52 cards. Find the probability of getting a king or a heart or a red card.*

Solution. Consider the following events :

A = getting a king, B = getting a heart

C = getting a red card.

NOTES

We have $P(A) = \frac{{}^{4}C_1}{{}^{52}C_1} = \frac{4}{52}$, $P(B) = \frac{{}^{13}C_1}{{}^{52}C_1} = \frac{13}{52}$

$$P(C) = \frac{{}^{26}C_1}{{}^{52}C_1} = \frac{26}{52}$$

$$P(A \cap B) = P(\text{getting a king of heart}) = \frac{1}{52}$$

$$P(B \cap C) = P(\text{getting a heart card}) = \frac{13}{52}$$

$$P(C \cap A) = P(\text{getting a red king}) = \frac{2}{52}$$

$$P(A \cap B \cap C) = P(\text{getting a king of heart}) = \frac{1}{52}$$

Required probability,

$P(A \cup B \cup C)$

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

$$= \frac{4}{52} + \frac{13}{52} + \frac{26}{52} - \frac{1}{52} - \frac{13}{52} - \frac{2}{52} + \frac{1}{52} = \frac{28}{52} = \frac{7}{13}.$$

Example 12. *Consider an experiment throwing a pair of dice. Let A and B be the events given by A = the sum of points is 8 ; B = there is an even number on first die. Find P(A/B) and P(B/A).*

Solution. We have $A = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$

and $B = \{(2, 1), ..., (2, 6), (4, 1), ..., (4, 6), (6, 1), ..., (6, 6)\}$

$\therefore$ $P(A) = \frac{5}{36}$ and $P(B) = \frac{18}{36}$

Now P(A/B) = Probability of occurrence of A when B occurs

= Probability of getting 8 as the sum when there is an even number on first die

$$= \frac{n(A \cap B)}{n(B)} = \frac{3}{18} = \frac{1}{6}$$

and P(B/A) = Probability of occurrence of B when A occurs

= Probability of getting an even number on first die when the sum of the numbers on two dice is 8

$$= \frac{n(A \cap B)}{n(A)} = \frac{3}{5}.$$

Example 13. *A bag contains 10 white and 15 black balls. Two balls are drawn in succession without replacement. What is the probability that first is white and second is black ?*

Solution. Consider the following events :

A = getting a white ball in first draw

B = getting a black ball in second draw.

Required probability = Probability of getting a white ball in first draw and black ball in second draw.

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

= P(A) P(B/A)

NOTES

Now $P(A) = \frac{^{10}C_1}{^{25}C_1} = \frac{10}{25} = \frac{2}{5}$

and P(B/A) = Probability of getting a black ball in second draw when a white ball has already been in first draw.

$= \frac{^{15}C_1}{^{24}C_1} = \frac{15}{24} = \frac{5}{8}$

($\because$ 24 balls are left after drawing a white ball in first draw out of which 15 are black)

So required probability = $P(A \cap B) = P(A)\ P(B/A)$

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

Example 14. *Two balls are drawn from an urn containing 2 white, 3 red and 4 black balls one by one without replacement. What is the probability that at least one ball is red ?*

Solution. Consider the following events :

A = not getting a red ball in first draw

B = not getting a red ball in second draw

Required probability = Probability that at least one ball is red

= 1 – Probability that none is red

= 1 – P(A and B) = $1 - P(A \cap B)$

= 1 – P(A) P(B/A)

Now P(A) = Probability of not getting a red ball in first draw

= Probability of getting an other colour (white or black) ball in first draw

$= \frac{6}{9} = \frac{2}{3}$

When another colour ball is drawn in first draw there are 5 other colour (white or black) balls and 3 red balls, out of which one other colour ball can be drawn in $^5C_1$ ways.

$\therefore$ $P(B/A) = \frac{5}{8}$

Required probability = $1 - P(A)\ P(B/A) = 1 - \frac{2}{3} \times \frac{5}{8} = \frac{7}{12}$.

Example 15. *If A and B are two events such that P(A) = 0.5, P(B) = 0.6 and $P(A \cup B) = 0.8$, find P(A/B) and P(B/A).*

Solution. We have $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$\therefore$ $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.5 + 0.6 - 0.8 = 0.3$

Now, $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.3}{0.6} = \frac{1}{2}$

and $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.3}{0.5} = \frac{3}{5}$.

Example 16. *A coin is tossed twice and the four possible outcomes are assumed to be equally likely. If A is the event, 'both head and tail have appeared', and B be the event, 'at most one tail is observed', find P(A), P(B), P(A/B) and P(B/A).*

Solution. Here, S = {HH, HT, TH, TT}, A = {HT, TH} and B = {HH, HT, TH}.

$\therefore$ $A \cap B = \{HT, TH\}$

Now, $P(A) = \frac{n(A)}{n(S)} = \frac{2}{4} = \frac{1}{2}$

$P(B) = \frac{n(B)}{n(S)} = \frac{3}{4}$ and $P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{4} = \frac{1}{2}$

$\therefore$ $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/2}{3/4} = \frac{2}{3}$ and $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/2}{1/2} = 1.$

Example 17. *A coin is tossed thrice and all eight outcomes are equally likely.*

A = 'The first throw results in head'

B = 'The last throw results in tail'

Prove that events A and B are independent.

Solution. Let S be the sample space, then

S = {HHH, HHT, THH, HTH, TTH, HTT, THT, TTT}

A = {HHH, HHT, HTH, HTT}, B = {HHT, HTT, THT, TTT}

$A \cap B = \{HHT, HTT\}$

$P(A) = \frac{n(A)}{n(S)} = \frac{4}{8} = \frac{1}{2}$, $P(B) = \frac{n(B)}{n(S)} = \frac{4}{8} = \frac{1}{2}$

$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{8} = \frac{1}{4}$

Clearly $P(A \cap B) = \frac{1}{4} = P(A)\,P(B)$

Hence, A and B are independent events.

Example 18. *Events A and B are independent. Find P(B) if P(A) = 0.35 and $P(A \cup B) = 0.6$.*

Solution. We have $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cup B) = P(A) + P(B) - P(A)\,P(B)$ ($\because$ A and B are independent)

$= P(A) + P(B)\,[1 - P(A)]$

$0.6 = 0.35 + P(B)\,(1 - 0.35)$

$0.25 = 0.65\,P(B)$

$P(B) = \frac{0.25}{0.65} = \frac{5}{13}.$

Example 19. *X can solve 90% of the problems given in a book and Y can solve 70%. What is the probability that at least one of them will solve the problem, selected at random from the book ?*

Solution. Let A and B be the events defined as follows :

A = X solves the problem, B = Y solves the problem

Clearly A and B are independent events such that

$P(A) = \frac{90}{100} = \frac{9}{10}$ and $P(B) = \frac{70}{100} = \frac{7}{10}$

Now required probability = $P(A \cup B)$

$= 1 - P(\overline{A})\,P(\overline{B})$ ($\because$ A and B are independent events)

$= 1 - \left(1 - \frac{9}{10}\right)\left(1 - \frac{7}{10}\right) = 1 - \frac{1}{10} \times \frac{3}{10} = 0.97$

NOTES

EXERCISE 1.1

1. Find the probability of getting a head in throw a coin.
2. Three unbiased coins are tossed, find the probability of getting
   (*i*) all heads (*ii*) two heads
   (*iii*) one head (*iv*) at least one head
   (*v*) at least two heads.
3. A bag contains 7 white, 6 red and 5 black balls. Two balls are drawn at random. Find the probability that they will both be white.
4. Four cards are drawn from a pack of cards. Find the probability that
   (*i*) all are diamonds (*ii*) there is one card of each suit, and
   (*iii*) there are two spades and two hearts.
5. Two dice are thrown simultaneously. Find the probability of getting
   (*i*) an even number as the sum (*ii*) the sum as a prime number
   (*iii*) a total of at least 10 (*iv*) a doublet of even number.
6. Tickets numbered from 1 to 20 are mixed up together and then a ticket is drawn at random. What is the probability that the ticket has a number which is a multiple of 3 or 7 ?
7. A bag contains 50 tickets numbered 1, 2, 3, ..., 50 of which five are drawn at random and arranged in ascending order of magnitude ($x_1 < x_2 < x_3 < x_4 < x_5$). Find the probability that $x_3 = 30$.
8. If A, B, C are mutually and exhaustive events, find P(B), if $\frac{1}{3}$ P(C) = $\frac{1}{2}$ P(A) = P(B).
9. If P(A) = a and P(B) = b, then show that P(A/B) ≥ $(a + b - 1)/b$.
10. Two cards are drawn from a pack of 52 cards. What is the probability that either both are red or both are kings ?
11. Given two mutually exclusive events A and B such that P(A) = $\frac{1}{2}$ and P(B) = $\frac{1}{3}$. Find P(A or B).
12. A die is thrown twice and the sum of numbers appearing is observed to be 6. What is the conditional probability that the number 4 has appeared at least once ?
13. A bag contains 19 tickets, numbered from 1 to 19. A ticket is drawn and then another ticket is drawn without replacement. Find the probability that both tickets will show even numbers.
14. If A and B are two events such that P(A) = 0.3, P(B) = 0.6 and P(B/A) = 0.5, find P(A/B) and P(A ∪ B).
15. A bag contains 3 red and 4 black balls and another bag has 4 red and 2 black balls. One bag is selected at random and from the selected bag a ball is drawn. Let A be the event that the first bag is selected, B be the event that the second bag is selected and C be the event that the ball drawn is red. Find P(A), P(B), P(C/A) and P(C/B).
16. If P(A) = 0.4, P(B) = p, P(A ∪ B) = 0.6 and A and B are given to be independent events, find the value of p.
17. A bag contains 5 white, 7 red and 4 black balls. If four balls are drawn one by one with replacement, what is the probability that none is white ?
18. Two dice are thrown. Find the probability of getting an odd number on the first die and a multiple of 3 on the other.
19. A problem of statistics is given to 3 students whose chances of solving it are $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$. What is the probability that the problem is solved ?

NOTES

Answers

1. $\frac{1}{2}$ **2.** (*i*) $\frac{1}{8}$ (*ii*) $\frac{3}{8}$ (*iii*) $\frac{3}{8}$ (*iv*) $\frac{7}{8}$ (*v*) $\frac{1}{2}$

3. $\frac{7}{51}$ **4.** (*i*) $\frac{11}{4165}$ (*ii*) $\frac{2197}{20825}$ (*iii*) $\frac{468}{20825}$

5. (*i*) $\frac{1}{2}$ (*ii*) $\frac{5}{12}$ (*iii*) $\frac{1}{6}$ (*iv*) $\frac{1}{12}$

6. $\frac{2}{5}$ **7.** $\frac{551}{15134}$ **8.** $\frac{1}{6}$ **10.** $\frac{55}{221}$

11. $\frac{5}{6}$ **12.** $\frac{2}{5}$ **13.** $\frac{4}{19}$ **14.** $\frac{1}{4}$; 0.75

15. $\frac{1}{2}; \frac{1}{2}; \frac{3}{7}; \frac{2}{3}$ **16.** $\frac{1}{3}$ **17.** $\left(\frac{11}{16}\right)^4$ **18.** $\frac{1}{6}$

19. $\frac{3}{4}$.

1.7. BAYES THEOREM

An event A can occur only if any one of the set of exhaustive and mutually exclusive events $B_1, B_2, ..., B_n$ occurs. The probabilities $P(B_1), P(B_2), ..., P(B_n)$ and the conditional probabilities $P(A/B_i)$, $i = 1, 2, 3, ..., n$ for an event A to occur are known.

Then the conditional probability $P(B_i/A)$ when A has already occurred is given by

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

$$= \frac{P(B_i)\,P(A/B_i)}{P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2) + ... + P(B_n)\,P(A/B_n)}$$

SOLVED EXAMPLES

Example 1. *Two boxes contain respectively 4 white and 2 black and 1 white and 3 black balls. One ball is transferred from the first box into the second and then one ball is drawn from the second. It turns out to be black. What is the probability that the transferred ball was white ?*

Solution. Let B_1 be the event that the transferred ball (ball drawn from the first box) is white and B_2 be the event that the transferred ball is black.

$$P(B_1) = \frac{4}{6} = \frac{2}{3},\ P(B_2) = \frac{2}{6} = \frac{1}{3}$$

Let A be the event that the ball drawn from the second box (after a ball is transferred from the first box to the second box) is black, then

$$P(A/B_1) = \frac{3}{5},\ P(A/B_2) = \frac{4}{5}$$

$P(B_1/A)$ = The probability that the ball transferred from the first box is white when the ball drawn from the second box is known to be black.

NOTES

$$P(B_1/A) = \frac{P(B_1)\,P(A/B_1)}{P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2)}$$

$$= \frac{\frac{2}{3}\times\frac{3}{5}}{\frac{2}{3}\times\frac{3}{5}+\frac{1}{3}\times\frac{4}{5}} = \frac{\frac{2}{5}}{\frac{2}{5}+\frac{4}{15}} = \frac{2}{5}\times\frac{3}{2} = \frac{3}{5}.$$

Example 2. *The chance that doctor X will diagnose disease Y correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. A patient of doctor X, who had disease Y, died. What is the chance that his disease was correctly diagnosed ?*

Solution. Let B_1 be the event that the diagnosis is correct and B_2 be the event that the diagnosis is incorrect. Let A be the event that the patient dies. Then

$$P(B_1) = \frac{60}{100} = 0.6,\ P(B_2) = 1 - P(B_1) = 1 - 0.6 = 0.4$$

$$P(A/B_1) = \frac{40}{100} = 0.4,\ P(A/B_2) = \frac{70}{100} = 0.7$$

$P(B_1/A)$ = Probability that a patient was correctly diagnosed, given that he had died.

$$P(B_1/A) = \frac{P(B_1)\,P(A/B_1)}{P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2)}$$

$$= \frac{0.6\times0.4}{0.6\times0.4 + 0.4\times0.7} = \frac{0.24}{0.24+0.28}$$

$$= \frac{0.24}{0.52} = 0.4615 \quad \text{or} \quad 46.15\%.$$

Example 3. *The contents of urns I, II and III are as follows :*

1 white, 2 black and 3 red balls,
2 white, 1 black and 1 red balls, and
4 white, 5 black and 3 red balls.

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

Solution. Let B_1, B_2 and B_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

$$P(B_1) = P(B_2) = P(B_3) = \frac{1}{3} \qquad (\because\ n = 3 \text{ urns}, m = 1)$$

$$P(A/B_1) = \frac{1\times3}{{}^6C_2} = \frac{1}{5},\ P(A/B_2) = \frac{2\times1}{{}^4C_2} = \frac{1}{3}$$

and $$P(A/B_3) = \frac{4\times3}{{}^{12}C_2} = \frac{2}{11}$$

$$P(B_1/A) = \frac{P(B_1)\,P(A/B_1)}{P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2) + P(B_3)\,P(A/B_3)}$$

$$= \frac{\frac{1}{3}\times\frac{1}{5}}{\frac{1}{3}\times\frac{1}{5}+\frac{1}{3}\times\frac{1}{3}+\frac{1}{3}\times\frac{2}{11}} = \frac{1}{5}\times\frac{165}{118} = \frac{33}{118}$$

$$P(B_2/A) = \frac{P(B_2)\,P(A/B_2)}{P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2) + P(B_3)\,P(A/B_3)}$$

$$= \frac{\frac{1}{3}\times\frac{1}{3}}{\frac{1}{3}\times\frac{1}{5}+\frac{1}{3}\times\frac{1}{3}+\frac{1}{3}\times\frac{2}{11}} = \frac{1}{3}\times\frac{165}{118} = \frac{55}{118}$$

$$P(B_3/A) = 1 - [P(B_1/A) + P(B_2/A)]$$

$$= 1-\left(\frac{33}{118}+\frac{55}{118}\right) = 1-\frac{88}{118} = \frac{30}{118}.$$

Example 4. *In a bolt factory machines X, Y, Z manufacture respectively 25%, 35% and 40% of the total. Of their output 5, 4 and 2 percent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machines X, Y and Z ?*

Solution. Let B_1, B_2, B_3 denote the events that a bolt selected at random is manufactured by the machines X, Y and Z respectively and let A denote the event of its being defective. Then we have

$$P(B_1) = \frac{25}{100} = 0.25,\ P(B_2) = \frac{35}{100} = 0.35,\ P(B_3) = \frac{40}{100} = 0.40$$

The probability of drawing a defective bolt manufactured by machine X is

$$P(A/B_1) = \frac{5}{100} = 0.05$$

Similarly, $P(A/B_2) = \frac{4}{100} = 0.04$, $P(A/B_3) = \frac{2}{100} = 0.02$

$P(B_1/A)$ = The probability that a defective bolt selected at random is manufactured by machine A

$$P(B_1/A) = \frac{P(B_1)\,P(A/B_1)}{P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2) + P(B_3)\,P(A/B_3)}$$

$$= \frac{0.25\times0.05}{0.25\times0.05+0.35\times0.04+0.40\times0.02} = \frac{125}{345} = \frac{25}{69}$$

$$P(B_2/A) = \frac{P(B_2)\,P(A/B_2)}{P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2) + P(B_3)\,P(A/B_3)}$$

$$= \frac{0.35\times0.04}{0.25\times0.05+0.35\times0.04+0.40\times0.02} = \frac{140}{345} = \frac{28}{69}$$

$$P(B_3/A) = 1 - [P(B_1/A) + P(B_2/A)]$$

$$= 1-\left(\frac{25}{69}+\frac{28}{69}\right) = 1-\frac{53}{69} = \frac{16}{69}.$$

EXERCISE 1.2

1. A doctor has taken a vaccine from either storage unit P (which contains 30 current and 10 outdated vaccines), or from unit Q (which contains 20 current and 20 outdated vaccines), or from unit R (which contains 10 current and 30 outdated vaccines), but he is twice as likely to have taken it from unit P as from unit Q and twice as likely to have taken it from unit Q as from unit R. If the vaccine selected is outdated, what is the probability that it came from unit P ?

NOTES

2. A factory has two machines. The empirical evidence has established that machines I and II produce 30% and 70% of the output respectively. It has also been established that 5% and 1% of the output produced by these machines respectively was defective. A defective item is drawn at random. What is the probability that the defective item was produced by machine II ?

3. A doctor has decided to prescribe two new drugs to 200 heart patients, as follows : 50 get drug A, 50 get drug B and 100 get both. Drug A reduces the probability of a heart attack by 35%, drug B reduces the probability by 20% and the two drugs, when taken together, work independently. The 200 patients were chosen so that each has an 80% chance of having a heart attack. If a randomly selected patient has a heart attack, what is the probability that the patient was given both drugs ?

4. In a class of 75 students, 15 were considered to be very intelligent, 45 as medium and the rest below average. The probability that a very intelligent student fail in a viva-voce examination is 0.005, the medium student failing has a probability 0.05, and the corresponding probability for a below average student is 0.15. If a student is known to have passed the viva-voce examination, what is the probability that he is below average?

5. Suppose that there is a chance for a newly constructed flyover to collapse whether the design is faulty or not . The chance that the design is faulty is 5%. The chance that the flyover collapses if the design is faulty is 95%, otherwise it is 30%. A flyover collapsed. What is the probability that it collapsed because of faulty design ?

Answers

1. 0.3636 **2.** 0.318 **3.** 0.4176

4. 0.18 **5.** 0.1428.

NOTES

2. PROBABILITY DISTRIBUTIONS

STRUCTURE

Frequency distributions can be classified under two categories :

(*i*) Observed Frequency Distributions

(*ii*) Theoretical or Expected Frequency Distributions

Observed frequency distributions are based on actual observations and experimentation. If certain hypothesis is assumed, it is sometimes possible to derive mathematically what the frequency distribution of certain universe should be. Such distributions are called Theoretical Distributions.

Here, we will deal with two types of probability distributions :

(*i*) Discrete Probability Distributions

(*ii*) Continuous Probability Distributions

Under the first type we will deal with

(*i*) Binomial Distribution

(*ii*) Poisson Distribution

Under the second type we will deal with Normal Distribution.

Discrete random variables represent count data such as the number of defectives in a sample of n items. Continuous random variables represent measured data such as heights, distances, temperatures in a given interval, etc.

A discrete random variable assumes each of its values with a certain probability. A table listing all possible values that discrete random variable can take along with the associated probabilities is called a Discrete Probability Distribution.

2.1. BINOMIAL DISTRIBUTION

NOTES

Binomial distribution was discovered by James Bernoulli in the year 1700.

Let a random experiment be performed repeatedly and let the occurrence of an event in any trial be called a success and its non-occurrence a failure. Consider a series of n independent trials. Let a random variable X denote the number of successes in these n trials. Let p be the probability of a success and $q = 1 - p$ that of a failure in a single trial. Let p be constant for each trial.

The probability of r successes in n trials in a specified order (say) SSSFFS ... FFSF (where S represents success and F failure) is given by

$$P(SSSFFS \ldots FFSF) = P(S)\,P(S)\,P(S)\,P(F)\,P(F)\,P(S) \ldots P(F)\,P(S)\,P(F)$$

$$= pppqqp \ldots qpq$$

$$= \underbrace{p \,.\, pp \ldots p}_{r \text{ factors}} \,.\, \underbrace{qqq \ldots qq}_{(n-r) \text{ factors}}$$

$$= p^r\, q^{n-r}$$

But r successes in n trials can occur in ${}^nC_r$ ways and the probability for each of these ways is $p^r\, q^{n-r}$. Hence, the probability of r successes in n trials is given by

$$P(X = r) = {}^nC_r\, p^r\, q^{n-r}, \text{ where } p + q = 1 \text{ and } r = 0, 1, 2, \ldots, n$$

The probability distribution of the number of successes so obtained is called the Binomial probability distribution and X is called the Binomial Variate.

Note. (*i*) P(X = *r*) is usually written as P(*r*).

(*ii*) n and p in the binomial distribution are called the parameters of the distribution.

(*iii*) Each trial has only two possible outcomes called success and failure .

(*iv*) There is a finite number of trials say n.

(*v*) All trials are identical, *i.e.,* p (and hence q) is constant in each trial.

(*vi*) The trials are independent of each other.

(*vii*) If n independent trials repeated N times then the expected frequency of r successes is N . P(*r*).

2.2. APPLICATIONS OF BINOMIAL DISTRIBUTION

This distribution is mainly applied in problems concerning

(*i*) Number of defectives in a sample from production line.

(*ii*) Estimation of reliability of systems.

(*iii*) Number of rounds fired from a gun hitting a target.

(*iv*) Radar detection.

2.3. RECURRENCE FORMULA FOR THE BINOMIAL DISTRIBUTION

We have $P(r) = {}^nC_r\, p^r\, q^{n-r}$

and $P(r+1) = {}^nC_{r+1}\, p^{r+1}\, q^{n-(r+1)}$

NOTES

$$\frac{P(r+1)}{P(r)} = \frac{{}^nC_{r+1}\, p^{r+1}\, q^{n-r-1}}{{}^nC_r\, p^r\, q^{n-r}}$$

$$= \frac{n!}{(r+1)!\,(n-r-1)!} \times \frac{r!\,(n-r)!}{n!} \times \frac{p^{r+1}\, q^{n-r-1}}{p^r\, q^{n-r}}$$

$$= \frac{r!\,(n-r)\,(n-r-1)!}{(r+1)\,r!\,(n-r-1)!} \times \frac{p}{q} = \frac{n-r}{r+1} \cdot \frac{p}{q}$$

or $$P(r+1) = \frac{n-r}{r+1} \cdot \frac{p}{q} P(r),$$

which is the required recurrence formula. Using this formula successively, we can find P(1), P(2), ..., if P(0) is known.

1.4. MEAN, VARIANCE AND STANDARD DEVIATION OF BINOMIAL DISTRIBUTION

For binomial distribution, we have $P(r) = {}^nC_r\, p^r\, q^{n-r}$

The mean (μ) is given by

$$\text{Mean } (\mu) = \sum_{r=0}^{n} r \,.\, P(r) = \sum_{r=0}^{n} r \,.\, {}^nC_r\, p^r\, q^{n-r}$$

$$= 0 + 1 \,.\, {}^nC_1\, p^1\, q^{n-1} + 2\, {}^nC_2\, p^2\, q^{n-2} + \ldots + r\, {}^nC_r\, p^r\, q^{n-r} + \ldots + \ldots + n \,.\, {}^nC_n\, p^n\, q^{n-n}$$

$$= np\, q^{n-1} + \frac{2n(n-1)}{2!} p^2\, q^{n-2} + \ldots + n \,.\, p^n$$

$$= n\, p[q^{n-1} + (n-1)\, p\, q^{n-2} + \ldots + p^{n-1}]$$

$$= np(q+p)^{n-1} = np \qquad (\because \quad p+q=1)$$

Hence, the mean of the binomial distribution is np.

The variance (σ^2) is given by

$$\text{Variance } (\sigma^2) = \sum_{r=0}^{n} r^2\, P(r) - \mu^2 = \sum_{r=0}^{n} [r + r(r-1)]\, P(r) - \mu^2$$

$$= \sum_{r=0}^{n} r\, P(r) + \sum_{r=0}^{n} r(r-1)\, P(r) - \mu^2$$

$$= \mu + \sum_{r=0}^{n} r(r-1)\, {}^nC_r\, p^r\, q^{n-r} - \mu^2$$

$$= \mu + [2 \,.\, 1 \,.\, {}^nC_2\, p^2\, q^{n-2} + 3 \,.\, 2 \,.\, {}^nC_3\, p^3\, q^{n-3} + \ldots + n(n-1) \,.\, {}^nC_n\, p^n] - \mu^2$$

$$= \mu + \left[2\frac{n(n-1)}{2!} p^2 q^{n-2} + 6\frac{n(n-1)\,(n-2)}{3!} p^3 q^{n-3} + \ldots + n(n-1)\, p^n\right] - \mu^2$$

$$= \mu + n(n-1)\, p^2\, [q^{n-2} + (n-2)\, pq^{n-3} + \ldots + p^{n-2}] - \mu^2$$

$$= \mu + n(n-1)\, p^2\, (q+p)^{n-2} - \mu^2$$

$$= \mu + n(n-1)\, p^2 - \mu^2 \qquad (\because \quad p+q=1)$$

$$= np + n(n-1)\, p^2 - n^2 p^2 \qquad (\because \quad \mu = np)$$

$$= np\, [1 + (n-1)\, p - np]$$

$$\sigma^2 = np(1-p) = npq$$

Hence, the variance of the binomial distribution is npq.

NOTES

The standard deviation (σ) is given by

Standard deviation $(\sigma) = \sqrt{npq}$

Hence, the standard deviation of the binomial distribution is $\sqrt{npq}$.

Note. (*i*) $\gamma_1 = \dfrac{q-p}{\sqrt{npq}} = \dfrac{1-2p}{\sqrt{npq}}$ gives the measure of skewness of the binomial distribution.

If $p < \dfrac{1}{2}$, skewness is positive, if $p > \dfrac{1}{2}$, skewness is negative and if $p = \dfrac{1}{2}$, skewness is zero.

(*ii*) $\beta_2 = 3 + \dfrac{1-6pq}{npq}$ gives a measure of the kurtosis of the binomial distribution.

SOLVED EXAMPLES

Example 1. *A die is thrown 6 times. If getting an even number is a success, what is the probability of :*

(*i*) *no success* (*ii*) *exactly 5 successes*

(*iii*) *at least 5 successes* (*iv*) *at most 5 successes.*

Solution. Here, S = {1, 2, 3, 4, 5, 6}. Let A denote 'getting an even number'.

$$A = \{2, 4, 6\}$$

$$p = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

$$q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2},\ n = 6$$

We know that $P(r) = {}^nC_r\, p^r\, q^{n-r}$

(*i*) P(no success) $= P(r = 0) = {}^6C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{6-0} = \left(\frac{1}{2}\right)^6$

(*ii*) P(exactly 5 successes) $= P(5) = {}^6C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{6-5} = 6 \times \left(\frac{1}{2}\right)^6 = \frac{3}{32}$

(*iii*) P(at least 5 successes) $= P(r \geq 5)$

$= P(5 \text{ successes or } 6 \text{ successes})$

$= P(5) + P(6)$

$= {}^6C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{6-5} + {}^6C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^{6-6}$

$= \frac{3}{32} + \frac{1}{64} = \frac{6+1}{64} = \frac{7}{64}$

(*iv*) P(at most 5 successes) $= P(r \leq 5)$

$= 1 - P(r > 5) = 1 - P(r = 6) = 1 - \frac{1}{64} = \frac{63}{64}$.

Example 2. *The items produced by a company contains 5% defective items. What is the probability of getting 2 defective items in a sample of 10 items ?*

Solution. Here, $p = \dfrac{5}{100} = \dfrac{1}{20}$, $n = 10$, $r = 2$

NOTES

$$q = 1 - p = 1 - \frac{1}{20} = \frac{19}{20}$$

We know that $P(r) = {}^{n}C_r\, p^r\, q^{n-r}$

P(2 defective items) = P(r = 2)

$$= {}^{10}C_2 \left(\frac{1}{20}\right)^2 \left(\frac{19}{20}\right)^{10-2} = \frac{10 \times 9}{2} \times \frac{(19)^8}{(20)^{10}} = \frac{45 \times (19)^8}{(20)^{10}}.$$

Example 3. *A pair of dice thrown 10 times. If getting a doublet (same number on both) is considered a success, find the probability of*

(*i*) *no success* (*ii*) *3 successes.*

Solution. A doublet can be obtained when a pair of dice is thrown in

(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), *i.e.*, (6, 6) 6 ways.

The two dice can be thrown in $6^2 = 36$ ways.

$$p = \text{P(getting doublet)} = \frac{6}{36} = \frac{1}{6}$$

$$q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6},\ n = 10$$

We know that $P(r) = {}^{n}C_r\, p^r\, q^{n-r}$

(*i*) P(no success) $= P(0) = {}^{10}C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{10-0} = 1 \times \left(\frac{5}{6}\right)^{10} = \left(\frac{5}{6}\right)^{10}$

(*ii*) P(3 successes) $= P(3) = {}^{10}C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{10-3} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} \times \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^7$

$$= \frac{120}{216} \times \left(\frac{5}{6}\right)^7 = \frac{5}{9} \times \left(\frac{5}{6}\right)^7.$$

Example 4. *Five cards are drawn successively with replacement from a well-shuffled pack of 52 cards. What is the probability that*

(*i*) *none is spade* (*ii*) *only 3 cards are spade ?*

Solution. $p = \text{P(spade card)} = \frac{13}{52} = \frac{1}{4}$

$$q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4},\ n = 5$$

We know that $P(r) = {}^{n}C_r\, p^r\, q^{n-r}$

(*i*) P(none is spade) $= P(0) = {}^{5}C_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{5-0} = 1 \times 1 \times \left(\frac{3}{4}\right)^5 = \frac{243}{1024}$

(*ii*) P(only 3 cards are spade) $= P(3) = {}^{5}C_3 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^{5-3}$

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

$$= \frac{10 \times 9}{4^5} = \frac{90}{1024} = \frac{45}{512}.$$

NOTES

Example 5. *If the probability of hitting a target is 10% and 10 shots are fired independently. What is the probability that the target will be hit at least once ?*

Solution. Here, $p = \frac{10}{100} = \frac{1}{10}$

$$q = 1 - p = 1 - \frac{1}{10} = \frac{9}{10},\ n = 10$$

We know that $P(r) = {}^nC_r\, p^r\, q^{n-r}$

P(target will be hit at least once)

$$= P(r \geq 1) = 1 - P(r < 1)$$
$$= 1 - P(r = 0)$$
$$= 1 - {}^{10}C_0 \left(\frac{1}{10}\right)^0 \left(\frac{9}{10}\right)^{10-0} = 1 - 1 \times 1 \times \left(\frac{9}{10}\right)^{10}$$
$$= 1 - 0.3487 = 0.6513.$$

Example 6. *A policeman fires 4 bullets on a dacoit. The probability that the dacoit will be killed by a bullet is 0.6. What is the probability that dacoit is still alive?*

Solution. Here, $p = 0.6,\ q = 1 - p = 1 - 0.6 = 0.4,\ n = 4$

We know that $P(r) = {}^nC_r\, p^r\, q^{n-r}$

P(dacoit is still alive) = P(not killed)

$$= P(r = 0) = {}^4C_0\ (0.6)^0\ (0.4)^{4-0}$$
$$= 1 \times 1 \times (0.4)^4 = 0.0256.$$

Example 7. *Find the parameters of the binomial distribution for which mean = 4 and variance = 3.*

Solution. We know that for a binomial distribution

$$\text{Mean} = np \text{ and variance} = npq$$

Here, $np = 4$ and $npq = 3$

We have $\frac{npq}{np} = \frac{3}{4} \Rightarrow q = \frac{3}{4}$

$$p = 1 - q = 1 - \frac{3}{4} = \frac{1}{4}$$

$$n \times \frac{1}{4} = 4 \Rightarrow n = 16.$$

Example 8. *Comment on the following statement. The mean of a binomial distribution is 3 and standard deviation is* $\sqrt{5}$.

Solution. We know that mean = np and standard deviation = $\sqrt{npq}$ for a binomial distribution.

Here, $np = 3$ and $\sqrt{npq} = \sqrt{5}$ or $npq = 5$

Now, $\frac{npq}{np} = \frac{5}{3} \Rightarrow q = \frac{5}{3} > 1,$

which is not possible, because probability cannot exceed 1.

Example 9. *Obtain the mean and variance of a binomial distribution for which P(X = 3) = 16 P(X = 7) and n = 10.*

Solution. $P(X = 3) = {}^{10}C_3\, p^3\, q^{10-3} = {}^{10}C_3\, p^3\, q^7$

$$P(X = 7) = {}^{10}C_7\, p^7\, q^{10-7} = {}^{10}C_7\ p^7\, q^3$$

NOTES

According to the given condition

$$^{10}C_3\ p^3\ q^7 = 16 \times\ ^{10}C_7\ p^7\ q^3$$

$$p^3\ q^7 = 16 \times p^7\ q^3 \qquad (\because\ ^{10}C_3 = {}^{10}C_7)$$

$$\Rightarrow \quad q^4 = 16p^4$$

$$\Rightarrow \quad q^4 = (2p)^4 \quad \Rightarrow \quad q = 2p$$

In a binomial distribution

$$p + q = 1 \quad \Rightarrow \quad p + 2p = 1 \quad \Rightarrow \quad p = \frac{1}{3}$$

$$q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$$

$$\text{Mean} = np = 10 \times \frac{1}{3} = \frac{10}{3}$$

$$\text{Variance} = npq = 10 \times \frac{1}{3} \times \frac{2}{3} = \frac{20}{9}.$$

Example 10. *The probability of a man hitting a target is* $\frac{1}{4}$. *How many times must he fire so that the probability of his hitting the target at least once is greater than* $\frac{2}{3}$ *?*

Solution. Let the man hits the target n times.

Here, $p = \frac{1}{4}$ and $q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}$

P(hitting the target at least once) = P($r \geq 1$)

$$= 1 - \text{P}(r < 1) = 1 - \text{P}(r = 0)$$

According to given condition

$$1 - \text{P}(r = 0) > \frac{2}{3}$$

$$1 - {}^{n}\text{C}_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{n-0} > \frac{2}{3}$$

$$1 - (0.75)^n > \frac{2}{3}, \quad i.e., \quad (0.75)^n < \frac{1}{3}$$

$$(0.75)^n < 0.3333$$

Taking log on both sides, we have

$$n \log (0.75) < \log (0.3333)$$

$$n(-0.1249) < -0.47712$$

$$n(0.1249) > 0.47712$$

$$n > \frac{0.47712}{0.1249} \quad i.e., \quad n > 3.82$$

$$n = 4.$$

Example 11. *Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or six ?*

Solution. Here, p = the probability of getting 5 or 6 with one die

$$= \frac{2}{6} = \frac{1}{3}$$

NOTES

$$q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3},\ n = 6,\ N = 729$$

The expected number of times at least three dice showing five or six

$= N \cdot P(r \geq 3)$

$= 729 \times [P(r = 3) + P(r = 4) + P(r = 5) + P(r = 6)]$

$$= 729 \times \left[{}^6C_3 \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^3 + {}^6C_4 \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2 + {}^6C_5 \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^1 + {}^6C_6 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^6 \right]$$

$$= \frac{729}{3^6}[160 + 60 + 12 + 1] = 233.$$

Example 12. *Out of 800 families with 5 children each, how many families would be expected to have*

(i) 3 boys *(ii)5 girls* *(iii) either 2 or 3 boys ?*

Assume equal probabilities for boys and girls.

Solution. Here, $n = 5$, $N = 800$

$$p = P(\text{a boy}) = \frac{1}{2},\ q = P(\text{a girl}) = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

(*i*) The probability of having 3 boys out of 5 children = $P(r = 3)$

$$= {}^5C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{5-3} = 10 \times \left(\frac{1}{2}\right)^5$$

$$= \frac{10}{32} = 0.3125$$

The expected number of families = $N \cdot P(r = 3) = 800 \times 0.3125 = 250$

(*ii*) The probability of having 5 girls out of 5 children = $P(r = 0)$

$$= {}^5C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{5-0} = 1 \times 1 \times \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.03125$$

The expected number of families

$= N \cdot P(r = 0) = 800 \times 0.03125 = 25.$

(*iii*) The probability of having 2 or 3 boys out of 5 children = $P(r = 2 \text{ or } r = 3)$

$= P(r = 2) + P(r = 3)$

$$= {}^5C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{5-2} + {}^5C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{5-3}$$

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

$$= 20 \times \left(\frac{1}{2}\right)^5 = \frac{20}{32} = 0.625$$

The expected number of families = $N. P(r = 2 \text{ or } r = 3)$

$= 800 \times 0.625 = 500.$

Example 13. *In sampling a large number of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 2000 such samples, how many would be expected to contain at least 3 defective parts.*

Solution. Here, μ = mean no. of defectives = 2

$\Rightarrow$ $np = 2,\ n = 20$

$$p = \frac{2}{20} = \frac{1}{10} = 0.1$$

$$q = 1 - p = 1 - 0.1 = 0.9,\ N = 2000$$

The probability of having at least 3 defectives in a sample of 20 parts = $P(r \geq 3)$

$$= 1 - P(r < 3)$$

$$= 1 - [P(r = 0) + P(r = 1) + P(r = 2)]$$

$$= 1 - [{}^{20}C_0(0.1)^0\ (0.9)^{20-0} + {}^{20}C_1(0.1)^1\ (0.9)^{20-1} + {}^{20}C_2(0.1)^2\ (0.9)^{20-2}]$$

$$= 1 - [0.1216 + 0.2702 + 0.2852] = 1 - 0.677 = 0.323$$

The expected number of samples = N . $P(r \geq 3)$

$$= 2000 \times 0.323 = 646.$$

Example 14. *Find the binomial distribution whose mean is 5 and variance is* $\frac{10}{3}$.

Solution. We know that mean = np and variance = npq

So $$np = 5 \quad \text{and} \quad npq = \frac{10}{3}$$

Now, $$\frac{npq}{np} = \frac{10/3}{5} \quad \Rightarrow \quad q = \frac{2}{3}$$

$$p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}.\ \text{But } np = 5 \quad \Rightarrow \quad n = 15$$

Hence, binomial distribution is

$$P(r) = {}^{15}C_r \left(\frac{1}{3}\right)^r \left(\frac{2}{3}\right)^{15-r}.$$

Example 15. *Four coins are tossed 160 times. The number of times r heads occur is given below.*

r	*0*	*1*	*2*	*3*	*4*
No. of times	*8*	*34*	*69*	*43*	*6*

Fit a binomial distribution to this data on the hypothesis that coins are unbiased.

Solution. The coins are unbiased so the probability of getting head is = $\frac{1}{2}$.

So $$p = \frac{1}{2},\ q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

Here, $$n = 4,\ N = 160$$

$$f(r) = \text{expected frequency} = N\ .\ P(r)$$

$$P(0) = {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{4-0} = 1 \times 1 \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

Using recurrence relation, we have

$$P(r+1) = \frac{n-r}{r+1} \cdot \frac{p}{q} P(r) \qquad (\because \quad p = q \text{ and } n = 4)$$

$$P(1) = \frac{4-0}{0+1} P(0) = 4 \times P(0) \qquad = 4 \times \frac{1}{16} = \frac{1}{4}$$

$$P(2) = \frac{4-1}{1+1} P(1) = \frac{3}{2} \times P(1) \qquad = \frac{3}{2} \times \frac{1}{4} = \frac{3}{8}$$

$$P(3) = \frac{4-2}{2+1} P(2) = \frac{2}{3} \times P(2) \qquad = \frac{2}{3} \times \frac{3}{8} = \frac{1}{4}$$

$$P(4) = \frac{4-3}{3+1} P(3) = \frac{1}{4} \times P(3) \qquad = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

NOTES

r	$P(r)$	$N \cdot P(r)$
0	$\frac{1}{16}$	$160 \times \frac{1}{16} = 10$
1	$\frac{1}{4}$	$160 \times \frac{1}{4} = 40$
2	$\frac{3}{8}$	$160 \times \frac{3}{8} = 60$
3	$\frac{1}{4}$	$160 \times \frac{1}{4} = 40$
4	$\frac{1}{16}$	$160 \times \frac{1}{16} = 10$

EXERCISE 2.1

1. A pair of dice thrown 6 times. If getting a total of 9 is considered a success. What is the probability of at least 5 successes.

2. A die is thrown 6 times. If getting an odd number is a success. Find the probability of

(*i*) no success (*ii*) 5 successes

(*iii*) at least 5 successes (*iv*) at most 5 successes.

3. A coin is tossed 5 times. What is the probability of getting at least 3 heads ?

4. Find the probability distribution of the number of heads observed when a coin is tossed 3 times.

5. If on an average one ship in every ten is wrecked, find the probability that out of 5 ships expected to arrive, 4 at least will arrive safely.

6. A pair of dice is thrown 4 times. What is the probability of getting doublets at least twice?

7. "The mean and variance of a binomial distribution are respectively 6 and 9". Is this statement correct ?

8. A student is given a true-false examination with 8 questions. If he gets 6 or more correct answers, he passes the examination. Given that he guesses the answer to each question, find the probability that he passes the examination.

9. In a box containing 60 bulbs, 6 are defective. What is the probability that out of a sample of 5 bulbs

(*i*) none is defective (*ii*) exactly 2 are defective ?

10. The sum of mean and variance of a binomial distribution is 15 and the sum of their squares is 117. Determine the distribution.

11. Out of 2000 families with 4 children each, how many would you expect to have

(*i*) at least one boy (*ii*) 2 boys

(*iii*) 1 or 2 girls (*iv*) no girls ?

Assume equal probabilities for boys and girls.

12. In a sampling a large number of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be expected to contain at least 3 defective.

13. Assuming that 20% of the population of a city are literate, so that the chance of an individual being literate is $\frac{1}{5}$ and assuming that 100 investigators each take 10 individuals to see whether they are literate. How many investigations would you expect to report 3 or less were literate ?

NOTES

Answers

1. $\frac{49}{96}$ **2.** (*i*) $\frac{1}{64}$ (*ii*) $\frac{3}{32}$ (*iii*) $\frac{7}{64}$ (*iv*) $\frac{63}{64}$

3. $\frac{1}{2}$

4.

r	0	1	2	3
$P(r)$	1/8	3/8	3/8	1/8

5. $\frac{7}{5}\left(\frac{9}{10}\right)^4$ **6.** $\frac{19}{44}$ **7.** No. **8.** $\frac{37}{256}$

9. (*i*) $\left(\frac{9}{10}\right)^5$ (*ii*) $\frac{729}{10000}$ **10.** ${}^{27}C_r \left(\frac{1}{3}\right)^r \left(\frac{2}{3}\right)^{27-r}$; $r = 0, 1, 2, \ldots, 27$

11. (*i*) 1875 (*ii*) 750 (*iii*) 1250 (*iv*) 125

12. 323 **13.** 88 (approx.).

2.5. POISSON DISTRIBUTION

Poisson distribution is a discrete probability distribution. Poisson distribution was discovered by the French mathematician, Simeon Denis Poisson in 1837.

In a random experiment, let p be the probability of the occurrence of an event and let n trials be made in such a way that

(*i*) p is very small, *i.e.*, $p \to 0$

(*ii*) n is very large, *i.e.*, $n \to \infty$

(*iii*) $np = \lambda$ (say) is finite.

Then the probability of occurrence of this event r times is given by the Poisson distribution as

$$P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}, \text{ where } r = 0, 1, 2, 3, \ldots$$

Note. (*i*) P(X = *r*) is usually written as P(*r*).

(*ii*) λ is called the parameter of the distribution.

(*iii*) The sum of the probabilities P(*r*) for $r = 0, 1, 2, 3, \ldots$ is 1,

Since P(0) + P(1) + P(2) + P(3) + ...

$$= e^{-\lambda}\frac{\lambda^0}{0!} + e^{-\lambda}\frac{\lambda^1}{1!} + e^{-\lambda}\frac{\lambda^2}{2!} + e^{-\lambda}\frac{\lambda^3}{3!} + \ldots$$

$$= e^{-\lambda} + \lambda\frac{e^{-\lambda}}{1!} + \lambda^2\frac{e^{-\lambda}}{2!} + \lambda^3\frac{e^{-\lambda}}{3!} + \ldots$$

$$= e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right] = e^{-\lambda} \cdot e^{\lambda} = 1$$

(*iv*) The events must be random and independent of each other.

(*v*) Events must be rare events.

(*vi*) If n independent trials repeated N times then the expected frequency of r successes is N . P(*r*).

2.6. APPLICATIONS OF POISSON DISTRIBUTION

NOTES

This distribution is mainly applied in problems concerning

(*i*) The demand for a product.

(*ii*) Typographical errors in a book.

(*iii*) The occurrence of accidents in a factory over a period of time.

(*iv*) The pattern of arrival of customers at a check-out counter.

(*v*) Number of air accidents in some time.

(*vi*) Number of deaths in a area by rare disease.

(*vii*) Number of fragments from a shell hitting a target.

(*viii*) Number of printing mistakes at each page of the book.

2.7. RECURRENCE FORMULA FOR THE POISSON DISTRIBUTION

We have $\quad P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$ and $P(r+1) = e^{-\lambda}\frac{\lambda^{r+1}}{(r+1)!}$

$$\frac{P(r+1)}{P(r)} = \frac{e^{-\lambda}\frac{\lambda^{r+1}}{(r+1)!}}{e^{-\lambda}\frac{\lambda^r}{r!}} = \frac{\lambda^{r+1}}{(r+1)!}\times\frac{r!}{\lambda^r} = \frac{\lambda}{(r+1)\,r!}\times r! = \frac{\lambda}{(r+1)}$$

$$P(r+1) = \frac{\lambda}{(r+1)}P(r),$$

which is the required recurrence formula. Using this formula successively, we can find P(1), P(2), ..., if P(0) is known.

2.8. MEAN, VARIANCE AND STANDARD DEVIATION OF POISSON DISTRIBUTION

For Poisson distribution, we have

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

The mean (μ) is given by

$$\text{Mean }(\mu) = \sum_{r=0}^{\infty} r\,.\,P(r) = \sum_{r=0}^{\infty} r\,.\,e^{-\lambda}\frac{\lambda^r}{r!}$$

$$= e^{-\lambda}\sum_{r=0}^{\infty} r\frac{\lambda^r}{r!} = e^{-\lambda}\left[0 + 1.\frac{\lambda^1}{1!} + 2.\frac{\lambda^2}{2!} + 3.\frac{\lambda^3}{3!} + \ldots\right]$$

$$= e^{-\lambda}\left[\lambda + \lambda^2 + \frac{\lambda^3}{2!} + \ldots\right]$$

$$= e^{-\lambda}\,\lambda\left[1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\ldots\right] = e^{-\lambda}\,\lambda\,e^{\lambda} = \lambda$$

Hence, the mean of the Poisson distribution is equal to the parameter λ.

NOTES

The variance (σ^2) is given by

$$\text{Variance } (\sigma^2) = \sum_{r=0}^{\infty} r^2.\,\mathrm{P}(r) - \mu^2 = \sum_{r=0}^{\infty} r^2.\,e^{-\lambda}\frac{\lambda^r}{r!} - \mu^2$$

$$= e^{-\lambda}\left[0 + 1^2\frac{\lambda^1}{1!} + 2^2\frac{\lambda^2}{2!} + 3^2\frac{\lambda^3}{3!} + 4^2\frac{\lambda^4}{4!} + \ldots\right] - \mu^2$$

$$= e^{-\lambda}.\lambda\left[1+\frac{2\lambda}{1!}+\frac{3\lambda^2}{2!}+\frac{4\lambda^3}{3!}+\ldots\right] - \mu^2$$

$$= e^{-\lambda}\,\lambda\left[1+\frac{(1+1)\,\lambda}{1!}+\frac{(1+2)\,\lambda^2}{2!}+\frac{(1+3)\,\lambda^3}{3!}+\ldots\right] - \mu^2$$

$$= e^{-\lambda}\,\lambda\left[\left(1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\ldots\right)+\left(\frac{\lambda}{1!}+\frac{2\lambda^2}{2!}+\frac{3\lambda^3}{3!}+\ldots\right)\right] - \mu^2$$

$$= e^{-\lambda}\,\lambda\left[e^{\lambda}+\lambda\left(1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\ldots\right)\right] - \mu^2$$

$$= e^{-\lambda}\,\lambda[e^{\lambda}+\lambda\,e^{\lambda}] - \lambda^2 = e^{-\lambda}\,\lambda\,e^{\lambda}\,(1+\lambda) - \lambda^2 = \lambda(1+\lambda) - \lambda^2 = \lambda$$

Hence, the variance of the Poisson distribution is also λ.

The standard deviation (σ) is given by standard deviation (σ) = $\sqrt{\lambda}$.

Note. The mean and the variance of the Poisson distribution is same.

SOLVED EXAMPLES

Example 1. *Using Poisson distribution, find the probability that the aces of spades will be drawn from a pack of well-shuffled cards at least once in 104 consecutive trials.* *(Given $e^{-2} = 0.1353$)*

Solution. $p = \frac{1}{52}$, $n = 104$

$$\lambda = np = 104 \times \frac{1}{52} = 2$$

We know that $\mathrm{P}(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

$$\mathrm{P}(\text{at least once}) = \mathrm{P}(r \geq 1) = 1 - \mathrm{P}(r < 1)$$

$$= 1 - \mathrm{P}(r = 0) = 1 - e^{-2}\frac{2^0}{0!}$$

$$= 1 - e^{-2} = 1 - 0.1353 = 0.8647.$$

Example 2. *Suppose a book of 585 pages contain 43 typographical mistakes. If these mistakes are randomly distributed throughout the book. What is the probability that 10 pages, selected at random will be free of mistakes ?* *(Given $e^{-0.735} = 0.4795$)*

NOTES

Solution. $p = \frac{43}{585} = 0.0735$, $n = 10$

$\lambda = np = 0.0735 \times 10 = 0.735$

We know that $P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

Required probability $= P(r = 0) = e^{-0.735}\frac{(0.735)^0}{0!}$

$= e^{-0.735} = 0.4795.$

Example 3. *A car hire firm has two cars, which it hires out day-by-day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used the proportion of days on which some demand is refused.* *(Given $e^{-1.5} = 0.2231$)*

Solution. The mean of the Poisson distribution is λ.

$\lambda = 1.5$

We know that $P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

The proportion of days on which neither car is used

= Probability of there being no demand for the car

$= P(r = 0) = e^{-1.5}\frac{(1.5)^0}{0!} = e^{-1.5} = 0.2231$

The proportion of days on which some demand is refused

= Probability for the number of demands to be more than two

$= P(r > 2) = 1 - P(r \le 2)$

$= 1 - [P(r = 0) + P(r = 1) + P(r = 2)]$

$= 1 - \left[e^{-1.5} + e^{-1.5}\frac{(1.5)}{1!} + e^{-1.5}\frac{(1.5)^2}{2!}\right]$

$= 1 - [1 + 1.5 + 1.125]\ e^{-1.5} = 1 - 3.625 \times 0.2231 = 0.1913.$

Example 4. *Find the probability that at most 5 defective components will be found in a lot of 200, if experience shows that 2% of such components are defective. Also find the probability of more than 5 defective components. (Given $e^{-4} = 0.018$).*

Solution. Here, $p = \frac{2}{100} = \frac{1}{50}$, $n = 200$

$\lambda = np = 200 \times \frac{1}{50} = 4$

We know that $P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

Probability that at most 5 defective components will be found $= P(r \le 5)$

$= P(r = 0) + P(r = 1) + P(r = 2) + P(r = 3) + P(r = 4) + P(r = 5)$

$= e^{-4}\frac{(4)^0}{0!} + e^{-4}\frac{(4)^1}{1!} + e^{-4}\frac{(4)^2}{2!} + e^{-4}\frac{(4)^3}{3!} + e^{-4}\frac{(4)^4}{4!} + e^{-4}\frac{(4)^5}{5!}$

$= e^{-4}\left[1 + 4 + \frac{16}{2} + \frac{64}{6} + \frac{256}{24} + \frac{1024}{120}\right]$

$= 0.018 \times 42.86 = 0.7715$

Probability of more than 5 defective components $= P(r > 5)$

$= 1 - P(r \le 5) = 1 - 0.7715 = 0.2285.$

NOTES

Example 5. *It is given that 2% of the electric bulbs manufactured by a company are defective. Using Poisson distribution, find the probability that a sample of 200 bulbs will contain*

(i) no defective bulb (ii) 2 defective bulbs

(iii) atmost 3 defective bulbs (iv) at least 3 defective bulbs. (Given $e^{-4} = 0.0183$).

Solution. Here, $p = \frac{2}{100} = \frac{1}{50}$, $n = 200$

$$\lambda = np = 200 \times \frac{1}{50} = 4$$

We know that $P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

(*i*) Probability of no defective bulb = $P(r = 0)$

$$= e^{-4}\frac{(4)^0}{0!} = e^{-4} = 0.0183$$

(*ii*) Probability of 2 defective bulbs = $P(r = 2)$

$$= e^{-4}\frac{(4)^2}{2!} = 0.0183 \times \frac{16}{2} = 0.1464$$

(*iii*) Probability of at most 3 defective bulbs = $P(r \le 3)$

$$= P(r = 0) + P(r = 1) + P(r = 2) + P(r = 3)$$

$$= e^{-4} + e^{-4}\frac{(4)^1}{1!} + e^{-4}\frac{(4)^2}{2!} + e^{-4}\frac{(4)^3}{3!}$$

$$= e^{-4}\left[1 + 4 + 8 + \frac{32}{3}\right] = 0.0183 \times \frac{71}{3} = 0.4331$$

(*iv*) Probability of at least 3 defective bulbs = $P(r \ge 3)$

$$= 1 - P(r < 3) = 1 - [P(r = 0) + P(r = 1) + P(r = 2)]$$

$$= 1 - [e^{-4} + 4e^{-4} + 8e^{-4}] = 1 - e^{-4} \times 13 = 1 - 0.0183 \times 13$$

$$= 1 - 0.2379 = 0.7621.$$

Example 6. *Assume that the chance of an individual coal-miner being killed in a mine accident during a year is $\frac{1}{1500}$. Use Poisson distribution to calculate the probability that in a mine employing 375 minors, there will be at least one total accident in a year. (Given $e^{-0.25} = 0.78$).*

Solution. Here, $p = \frac{1}{1500}$, $n = 375$

$$\lambda = np = 375 \times \frac{1}{1500} = \frac{1}{4} = 0.25$$

We know that $P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

Probability of at least one total accident = $P(r \ge 1)$

$$= 1 - P(r < 1) = 1 - P(r = 0)$$

$$= 1 - e^{-0.25}\frac{(0.25)^0}{0!} = 1 - e^{-0.25} = 1 - 0.78 = 0.22.$$

Example 7. *A manufacturer knows that the razor blades he makes contain on the average 0.5% defectives. He packs them in packets of 5. What is the probability that a packet picked at random contains 3 or more defective blades ? (Given $e^{-0.025} = 0.9753$).*

NOTES

Solution. Here, $p = 0.5\% = \frac{0.5}{100} = 0.005$, $n = 5$

$\lambda = np = 5 \times 0.005 = 0.025$

We know that $P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

Probability of 3 or more defective blades = $P(r \geq 3)$

$= 1 - P(r < 3) = 1 - [P(r = 0) + P(r = 1) + P(r = 2)]$

$= 1 - \left[e^{-0.025} + e^{-0.025}\frac{(0.025)^1}{1!} + e^{-0.025}\frac{(0.025)^2}{2!}\right]$

$= 1 - e^{-0.025}[1 + 0.025 + 0.0003125]$

$= 1 - 0.9753 \times 1.0253 = 1 - 0.999975 = 0.00002491.$

Example 8. *If the variance of the Poisson distribution is 2, find probabilities for r = 1, 2, 3, 4 using recurrence relation of the Poisson distribution. Also find P(r ≥ 4). (Given $e^{-2} = 0.1353$).*

Solution. Here, variance = 2

So $\lambda = 2$

We know that $P(r) = e^{-\lambda}\frac{\lambda^r}{r!}$

$$P(0) = e^{-2}\frac{(2)^0}{0!} = e^{-2} = 0.1353$$

We know that the recurrence relation is

$$P(r + 1) = \frac{\lambda}{r+1}P(r)$$

Now putting $r = 0, 1, 2, 3$ in the recurrence relation, we have

$$P(1) = \frac{\lambda}{1}P(0) = 2 \times 0.1353 = 0.2706$$

$$P(2) = \frac{\lambda}{2}P(1) = \frac{2}{2} \times 0.2706 = 0.2706$$

$$P(3) = \frac{\lambda}{3}P(2) = \frac{2}{3} \times 0.2706 = 0.1804$$

$$P(4) = \frac{\lambda}{4}P(3) = \frac{2}{4} \times 0.1804 = 0.0902$$

and $P(r \geq 4) = 1 - P(r < 4)$

$= 1 - [P(r = 0) + P(r = 1) + P(r = 2) + P(r = 3)]$

$= 1 - [P(0) + P(1) + P(2) + P(3)]$

$= 1 - (0.1353 + 0.2706 + 0.2706 + 0.1804)$

$= 1 - 0.8569 = 0.1431.$

Example 9. *A manufacturer who produces medicine bottles finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 1000 boxes from the producer of bottles. Using Poisson distribution, find how many boxes will contain :*

(i) no defective bottle

(ii) at least two defective bottles. *(Given $e^{-0.5} = 0.6065$)*

Solution. Here, $p = 0.1\% = \frac{0.1}{100} = 0.001$, $n = 500$

$\lambda = np = 500 \times 0.001 = 0.5$, N = 1000

NOTES

Number of boxes containing no defective bottle = N . P(r = 0)

$$= 1000 \times e^{-0.5} \frac{(0.5)^0}{0!}$$

$$= 1000 \times 0.6065 = 606.5 = 606 \text{ (approx.)}$$

Number of boxes containing at least two defective bottles = N . P($r \geq 2$)

$$= N \cdot [1 - P(r < 2)]$$

$$= N \times [1 - (P(r = 0) + P(r = 1))]$$

$$= 1000 \times \left[1 - \left(e^{-0.5} + e^{-0.5} \frac{(0.5)^1}{1}\right)\right]$$

$$= 1000 \times [1 - (0.6065 \times 1.5)] = 1000 \times [1 - 0.90975]$$

$$= 1000 \times 0.09025 = 90.25 = 90 \text{ (approx.)}.$$

Example 10. *After correcting 100 pages of a book, the proof-reader finds that there are on the average, 4 errors per 10 pages. How many pages would one expect to find with 0, 1 and 2 errors in 1000 pages of the first print of the book ?*

(Given $e^{-0.4} = 0.6703$).

Solution. Here, λ = average number of errors per page

$$= \frac{4}{10} = 0.4, \text{ N} = 1000$$

We know that $P(r) = e^{-\lambda} \frac{\lambda^r}{r!}$

(*i*) Probability of no errors = P(r = 0) = $e^{-0.4} \frac{(0.4)^0}{0!} = e^{-0.4} = 0.6703$

Number of pages containing no errors = N . P(r = 0)

$$= 1000 \times 0.6703 = 670.3 = 670 \text{ (approx.)}$$

(*ii*) Probability of one error = P(r = 1)

$$= e^{-0.4} \frac{(0.4)^1}{1!} = 0.6703 \times 0.4 = 0.26812$$

Number of pages containing one error = N . P(r = 1)

$$= 1000 \times 0.26812 = 268.12 = 268 \text{ (approx.)}$$

(*iii*) Probability of two errors = P(r = 2)

$$= e^{-0.4} \frac{(0.4)^2}{2!} = 0.6703 \times 0.08 = 0.053624$$

Number of pages containing two errors = N . P(r = 2)

$$= 1000 \times 0.053624 = 53.624 = 54 \text{ (approx.)}.$$

Example 11. *For a Poisson variate X, calculate P(X > 0), if it is given that*

4P(X = 4) = 5P(X = 5).

Solution. Given 4P(X = 4) = 5P(X = 5)

$$4 \cdot e^{-\lambda} \frac{\lambda^4}{4!} = 5 \cdot e^{-\lambda} \frac{\lambda^5}{5!}$$

$$4 \frac{\lambda^4}{4!} = 5 \frac{\lambda^5}{5 \times 4!}$$

NOTES

$$4\lambda^4 = \lambda^5 \quad \Rightarrow \quad \lambda = 4$$

$$P(X > 0) = P(r > 0) = 1 - P(r \leq 0) = 1 - P(r = 0)$$

$$= 1 - e^{-4}\frac{(4)^0}{0!} = 1 - e^{-4} = 1 - 0.0183 = 0.9817.$$

Example 12. *The frequency of accidents per shift in a factory is shown in the following data :*

Accidents per shift	*Frequency*
0	*192*
1	*100*
2	*24*
3	*3*
4	*1*
Total	*320*

Calculate the mean number of accidents per shift. Fit a Poisson distribution and calculate theoretical frequencies.

Solution. Mean number of accidents per shift

$$= \frac{\Sigma fx}{\Sigma f} = \frac{0 + 100 + 48 + 9 + 4}{320} = \frac{161}{320} = 0.5031$$

$$\lambda = 0.5031$$

Required Poisson distribution $= N \cdot e^{-\lambda}\frac{\lambda^r}{r!}$

$$= 320 \times e^{-0.5031} \times \frac{(0.5031)^r}{r!}$$

$$= \frac{(193.48)\,(0.5031)^r}{r!}$$

r	*N.P(r)*	*Theoretical frequencies*
0	193.48	193
1	97.34	97
2	24.49	24
3	4.10	4
4	0.51	1

Example 13. *A typist kept a record of mistakes per day during 300 working days :*

Mistakes per day	*0*	*1*	*2*	*3*	*4*
Number of days	*143*	*90*	*44*	*14*	*9*

Fit a Poisson distribution for the above data and calculate theoretical frequencies.

NOTES

Solution. Here, $\lambda = \text{mean} = \dfrac{\Sigma fx}{\Sigma f}$

$$= \frac{0 + 90 + 88 + 42 + 36}{300} = \frac{256}{300} = 0.853,\ N = 300$$

$$P(r) = e^{-0.853}\frac{(0.853)^r}{r!} = (0.426)\frac{(0.853)^r}{r!}$$

$$P(0) = (0.426)\frac{(0.853)^0}{0!} = 0.426$$

$$P(1) = (0.426)\frac{(0.853)^1}{1!} = 0.426 \times 0.853 = 0.363$$

$$P(2) = (0.426)\frac{(0.853)^2}{2!} = 0.426 \times 0.3638 = 0.155$$

$$P(3) = (0.426)\frac{(0.853)^3}{3!} = 0.426 \times 0.1034 = 0.044$$

$$P(4) = (0.426)\frac{(0.853)^4}{4!} = 0.426 \times 0.0221 = 0.009$$

r	*N . P(r)*	*Theoretical frequencies*
0	127.8	128
1	108.9	109
2	46.5	47
3	13.2	13
4	2.7	3

EXERCISE 2.2

1. Suppose a book of 600 pages contain 40 printing mistakes. If these mistakes are randomly distributed throughout the book. What is the probability that 10 pages, selected at random, will be free of mistakes ? (Given $e^{-0.67} = 0.51$)

2. Suppose 300 misprints are distributed randomly throughout a book of 500 pages. Find the probability that a given page contains (*i*) exactly 2 misprints (*ii*) 2 or more misprints.

3. Suppose 2 percent of the items made by a factory are defective. Find the probability that there are 3 defective items in a sample of 100 items. (Given $e^{-2} = 0.135$)

4. If the probability that an individual suffers a bad reaction from a certain injection is 0.001, determine the probability that out of 2000 individuals

(*i*) exactly 3 (*ii*) more than 2 individuals

(*iii*) none (*iv*) more than 1 individual, will suffer a bad reaction.

5. An insurance company finds that 0.005% of the population dies from a certain kind of accident each year. What is the probability that the company must pay off no more than 3 of 10,000 insured risks against such accident in a given year ? (Given $e^{-0.5} = 0.6065$)

6. A manufacturer of screws knows that 4% of his product is defective. If he sells the screws in boxes of 100 and guarantee that not more than 5 screws will be defective. What is the probability that a box will fail to meet the guaranteed quality ?

7. A manufacturer knows that the condensers he makes contain on the average 1% of the defectives. He packs them in boxes of 100. What is the probability that a box picked at random will contain 4 or more defective condensers ?

NOTES

8. Assume that the probability of an individual coal-miner being killed in a mine accident during a year is $\frac{1}{2400}$. Use Poisson distribution to calculate the probability that in a mine employing 200 miners, there will be at least one fatal accident in a year.

9. An insurance company found that only 0.01% of the population is involved in a certain type of accident each year. If its 1000 policy holders were randomly selected from the population, then what is the probability that not more than two of its clients are involved in such an accident next year ? (Given $e^{-0.1} = 0.9048$)

10. If X is a Poisson variate such that $P(X = 2) = 9\ P(X = 4) + 90\ P(X = 6)$, find the mean of X.

11. If X is a Poisson variate such that $P(X = 1) = 0.01487$; $P(X = 2) = 0.04461$, find $P(X = 3)$.

12. If X is a Poisson variate such that $P(X = 1) = P(X = 2)$; find

(*i*) mean of the distribution (*ii*) $P(X = 0)$ (*iii*) $P(X = 4)$

(Given $e^{-2} = 0.1353$)

13. The number of accidents in a year involving taxi drivers in a city follows a Poisson distribution with mean equal to 3. Out of 1000 taxi drivers, find approximately the number of drivers with

(*i*) no accident in a year (*ii*) more than 3 accident in a year.

14. In a certain factory turning out razor blades, there is small chance $\frac{1}{500}$ for any blade to be defective. The blades are supplied in packets of 10. Using Poisson's distribution, calculate the approximate number of packets containing

(*i*) no defective

(*ii*) one defective and

(*iii*) two defective blades respectively in a consignment of 10,000 packets.

(Given $e^{-0.02} = 0.9802$)

15. The distribution of typing mistakes committed by a typist is given below :

Mistakes per page	0	1	2	3	4	5
No. of pages	142	156	69	27	5	1

Assuming Poisson model, find out the expected frequencies.

16. Accidents per day were recorded in a certain city for a period of 400 days, as follows :

No. of accidents	0	1	2	3	4	5
No. of days	213	128	37	18	3	1

Assuming Poisson model, find out the expected frequencies.

17. The first proof of 200 pages of a book containing 560 pages revealed the following distribution of the number of printing errors :

No. of errors in a page	0	1	2	3	4	5
No. of pages	112	63	20	3	1	1

Fit a Poisson distribution corresponding to these data.

Answers

1. 0.51 **2.** (*i*) 0.1 (*ii*) 0.122

3. 0.18

4. (*i*) 0.18 (*ii*) 0.325 (*iii*) 0.135 (*iv*) 0.59

NOTES

5. 0.3235 **6.** $1 - e^{-4}\left[5 + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!} + \frac{4^5}{5!}\right]$

7. 0.019 **8.** 0.08 **9.** 0.9998

10. 1 **11.** 0.08922

12. (*i*) 2 (*ii*) 0.1353 (*iii*) 0.0902

13. (*i*) 50 (*ii*) 353

14. (*i*) 9802 (*ii*) 196 (*iii*) 2

15. 147, 147, 74, 25, 6, 1 pages **16.** 202, 138, 47, 11, 2, 0

17. 109, 66, 20, 4, 1, 0

2.9. NORMAL DISTRIBUTION

The normal distribution was discovered by French mathematician De-Moivre in 1733. It was derived from the binomial distribution in the limiting case. The normal distribution is a continuous distribution.

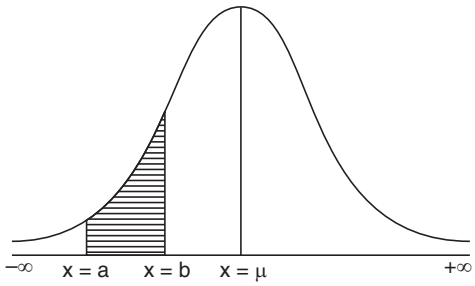
A random variable X is said to have a normal distribution with mean μ and standard deviation σ if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \; ; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0,$$

where $e = 2.7183$ and $\sqrt{2\pi} = 2.5066$.

The normal distribution with mean μ and variance σ^2 is denoted by $N(\mu, \sigma^2)$.

The graph of normal distribution is called the normal curve. It is bell-shaped and symmetrical about mean μ. The two tails of the curve extend to $+\infty$ and $-\infty$ towards the positive and negative directions of the *x*-axis respectively and gradually approach the *x*-axis without ever meeting it.



For a normally distributed random variable *x* with mean μ and variance σ^2 the probability that *x* lies between *a* and *b* is given by

$P(a < x < b)$ = area under the normal curve $f(x)$ between

$x = a$ and $x = b$.

2.10. PROPERTIES OF THE NORMAL DISTRIBUTION

The normal probability curve with mean μ and standard deviation σ is given by the equation

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

NOTES

and has the following properties :

(*i*) $f(x) \geq 0$

(*ii*) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ *i.e.*, the total area under the normal curve above the x-axis is 1.

(*iii*) The normal curve is bell-shaped and symmetrical about the line $x = \mu$, *i.e.,* mean.

(*iv*) It is a unimodal distribution *i.e.*, mean = median = mode.

(*v*) The height of the normal curve is maximum at the mean value. The maximum ordinate at $x = \mu$ is given by $y = \dfrac{1}{\sigma\sqrt{2\pi}}$.

(*vi*) $P(\mu - \sigma < x < \mu + \sigma) = 68\%$

$P(\mu - 2\sigma < x < \mu + 2\sigma) = 95.5\%$

$P(\mu - 3\sigma < x < \mu + 3\sigma) = 99.7\%$.

2.11. STANDARD FORM OF THE NORMAL DISTRIBUTION

A random variable Z which has a normal distribution with $\mu = 0$ and $\sigma = 1$ is said to have a standard distribution. The probability density function for the normal distribution in standard form is given by

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

where $Z = \dfrac{x - \mu}{\sigma}$; Z is called the standardized normal random variable and is denoted by N(0, 1).

$P(a \leq Z \leq b)$ = area under the standard normal curve between $Z = a$ and $Z = b$.

Note. The probabilities $P(z_1 \leq Z \leq z_2)$, $P(z_1 < Z \leq z_2)$, $P(z_1 \leq Z < z_2)$ and $P(z_1 < Z < z_2)$ are all regarded to be the same.

SOLVED EXAMPLES

Example 1. *The marks obtained by a group of students who appeared for a test were normally distributed with mean 80 and standard deviation 6. Find the standard scores for the student who scored*

(*i*) *98 marks* (*ii*) *58 marks* (*iii*) *50 marks.*

Solution. Suppose x is normally distributed with mean (μ) = 80 and standard deviation (σ) = 6.

We know that

Standard normal variate $Z = \dfrac{x - \mu}{\sigma}$

(*i*) When $x = 98$, $Z = \dfrac{98 - 80}{6} = \dfrac{18}{6} = 3$

(*ii*) When $x = 58$, $Z = \dfrac{58 - 80}{6} = \dfrac{-22}{6} = -3.67$

(*iii*) When $x = 50$, $Z = \dfrac{50 - 80}{6} = \dfrac{-30}{6} = -5$.

NOTES

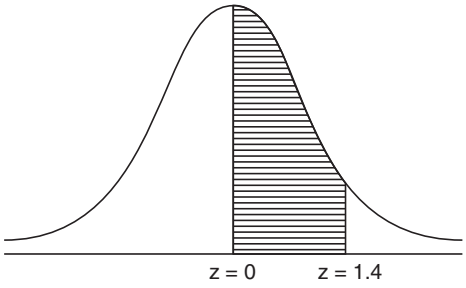
Example 2. *Find the area under the standard normal curve in each of the following :*

(*i*) *P(0 ≤ Z ≤ 1.4)* (*ii*) *P(– 1.67 ≤ Z < 0)*

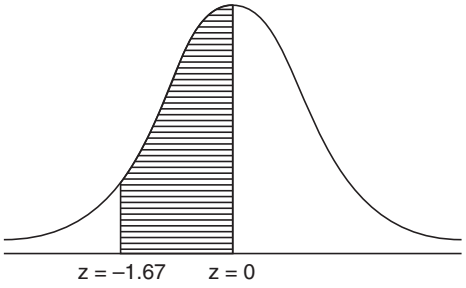
(*iii*) *P(0.65 ≤ Z < 2.35)* (*iv*) *P(– 3 ≤ Z < 1.6).*

Solution. Using the table of the area of standard normal curve, we have

(*i*) P(0 ≤ Z ≤ 1.4) = 0.4192



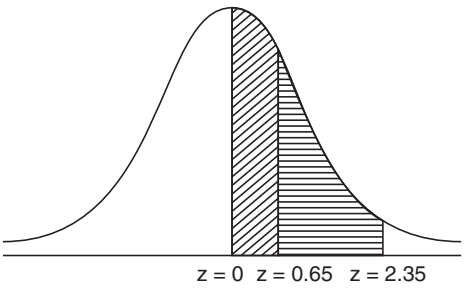
(*i*)



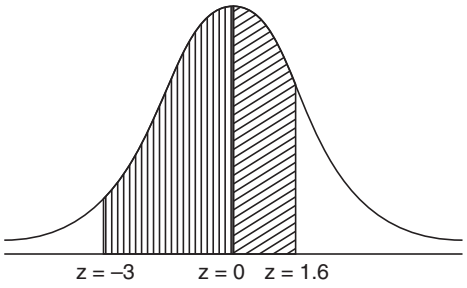
(*ii*)

(*ii*) P(– 1.67 ≤ Z < 0) = P(0 < Z ≤ 1.67) = 0.4525
(by symmetry)

(*iii*) P(0.65 ≤ Z < 2.35) = P(0 < Z ≤ 2.35)
– P(0 < Z ≤ 0.65)
= 0.4906 – 0.2422
= 0.2484.



(*iv*) P(– 3 ≤ Z < 1.6) = P(– 3 ≤ Z < 0)
+ P(0 ≤ Z < 1.6)
= P(0 < Z ≤ 3) + P(0 ≤ Z < 1.6)
(by symmetry)
= 0.4987 + 0.4452
= 0.9439.



Example 3. *Students of a class were given an aptitude test. Their marks were found to be normally distributed with mean 60 and standard deviation 5. What percentage of students scored more than 60 marks ?*

Solution. Here, $\mu = 60$, $\sigma = 5$, $x = 60$

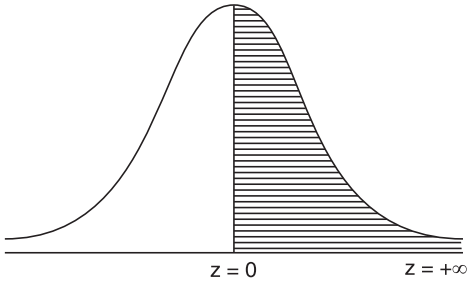
$$Z = \frac{x - \mu}{\sigma}$$

$$= \frac{60 - 60}{5} = 0$$

$P(x > 60) = P(Z > 0)$

$= P(0 < Z < \infty)$

$= 0.5 = 50\%$.



NOTES

Example 4. *A sample of 100 dry battery cells tested to find the length of life produced the following results :*

μ = *12 hours,* σ = *3 hours*

Assuming data to be normally distributed, what percentage of battery cells are expected to have life

(*i*) *more than 15 hours* (*ii*) *less than 6 hours*

(*iii*) *between 10 and 14 hours ?*

Solution. Here, x denotes the length of life of dry battery cells.

We know that $Z = \frac{x - \mu}{\sigma} = \frac{x - 12}{3}$

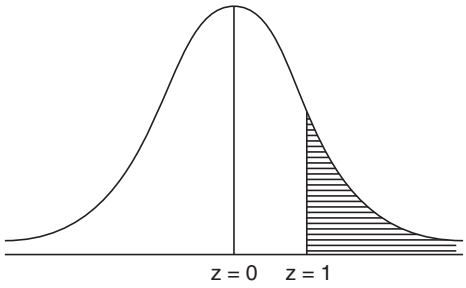
(*i*) When $x = 15$, $Z = \frac{15 - 12}{3} = \frac{3}{3} = 1$

$P(x > 15) = P(Z > 1)$

$= P(0 < Z < \infty) - P(0 < Z < 1)$

$= 0.5 - 0.3413$

$= 0.1587 = 15.87\%$.



(*ii*) When $x = 6$, $Z = \frac{6 - 12}{3} = \frac{-6}{3} = -2$

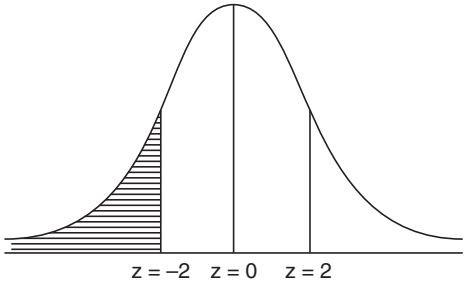
$P(x < 6) = P(Z < -2)$

$= P(Z > 2)$

$= P(0 < Z < \infty) - P(0 < Z < 2)$

$= 0.5 - 0.4772$

$= 0.0228 = 2.28\%$



(*iii*) When $x = 10$,

$Z = \frac{10 - 12}{3} = \frac{-2}{3} = -0.67$

When $x = 14$, $Z = \frac{14 - 12}{3} = \frac{2}{3} = 0.67$

$P(10 < x < 14) = P(-0.67 < Z < 0.67)$

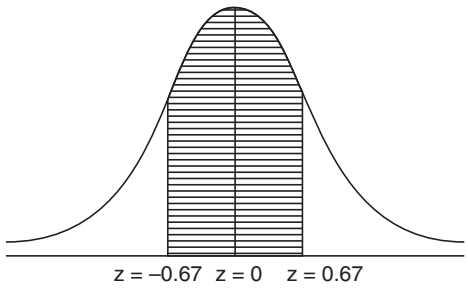
$= P(-0.67 < Z < 0) + P(0 < Z < 0.67)$

$= P(0 < Z < 0.67) + P(0 < Z < 0.67$

$= 2P(0 < Z < 0.67)$

$= 2 \times 0.2486 = 0.4972$

$= 49.72\%$.



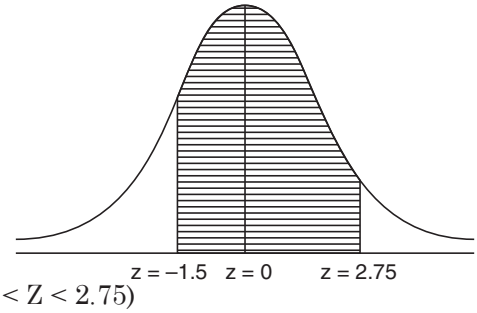
Example 5. *A normal distribution is given with mean 50 and standard deviation 8. Find the probability that x assumes a value between 38 and 72.*

Solution. Here, $\mu = 50$, $\sigma = 8$

$Z = \frac{x - \mu}{\sigma} = \frac{x - 50}{8}$

When $x = 38$, $Z = \frac{38 - 50}{8} = \frac{-12}{8} = -1.5$

When $x = 72$, $Z = \frac{72 - 50}{8} = \frac{22}{8} = 2.75$



$P(38 < x < 72) = P(-1.5 < Z < 2.75)$

$= P(-1.5 < Z < 0) + P(0 < Z < 2.75)$

$= P(0 < Z < 1.5) + P(0 < Z < 2.75)$

$= 0.4332 + 0.4970 = 0.9302$.

Example 6. *In a sample of 1000 cases, the mean of a certain test is 14 and standard deviation is 2.5. Assuming the distribution to be normal, find :*

(*i*) *how many students score between 12 and 15 ?*

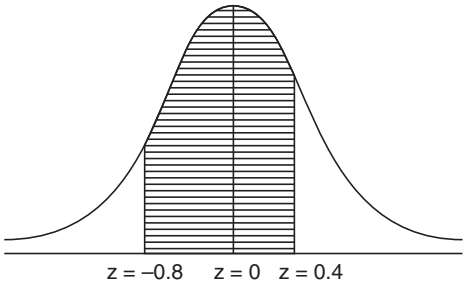
(*ii*) *how many scores above 18 ?*

(*iii*) *how many scores below 8 ?*

(*iv*) *how many scores 16 ?*

Solution. Here, N = 1000, $\mu = 14$ and $\sigma = 2.5$

$$Z = \frac{x - \mu}{\sigma} = \frac{x - 14}{2.5}$$



(*i*) When $x = 12$,

$$Z = \frac{12 - 14}{2.5} = \frac{-2}{2.5} = -0.8$$

When $x = 15$,

$$Z = \frac{15 - 14}{2.5} = \frac{1}{2.5} = 0.4$$

$$P(12 < x < 15) = P(-0.8 < Z < 0.4)$$

$$= P(-0.8 < Z < 0) + P(0 < Z < 0.4)$$

$$= P(0 < Z < 0.8) + P(0 < Z < 0.4)$$

$$= 0.2881 + 0.1554 = 0.4435$$

Number of students scoring between 12 and 15 $= 1000 \times 0.4435$

$= 443.5 \simeq 444$ (approx.)

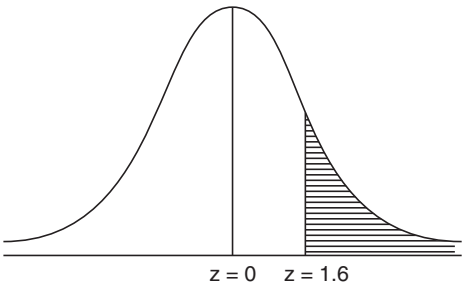
(*ii*) When $x = 18$, $Z = \frac{18 - 14}{2.5} = \frac{4}{2.5} = 1.6$

$$P(x > 18) = P(Z > 1.6)$$

$$= P(0 < Z < \infty) - P(0 < Z < 1.6)$$

$$= 0.5 - 0.4452$$

$$= 0.0548$$



Number of students scoring above 18

$= 1000 \times 0.0548$

$= 54.8 \simeq 55$ (approx.)

(*iii*) When $x = 8$, $Z = \frac{8 - 14}{2.5} = \frac{-6}{2.5} = -2.4$

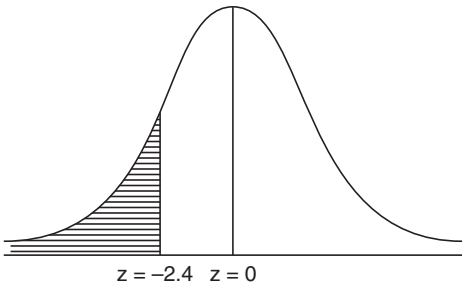
$$P(x < 8) = P(Z < -2.4)$$

$$= P(Z > 2.4)$$

$$= P(0 < Z < \infty) - P(0 < Z < 2.4)$$

$$= 0.5 - 0.4918$$

$$= 0.0082$$



Number of students scoring below 8

$= 1000 \times 0.0082$

$= 8.2 \simeq 8$ (approx.)

(*iv*) Area between $x = 15.5$ and $x = 16.5$

When $x = 15.5$, $Z = \frac{15.5 - 14}{2.5} = \frac{1.5}{2.5} = 0.6$

NOTES

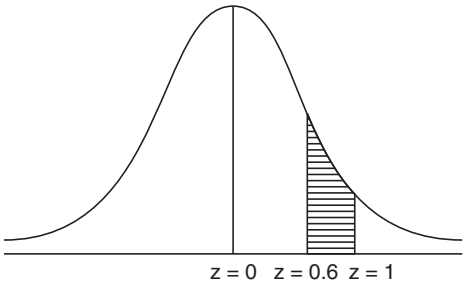
When $x = 16.5$, $Z = \frac{16.5 - 14}{2.5} = \frac{2.5}{2.5} = 1$

$P(15.5 < x < 16.5) = P(0.6 < Z < 1)$

$= P(0 < Z < 1) - P(0 < Z < 0.6)$

$= 0.3413 - 0.2257$

$= 0.1156$



Number of students scoring 16

$= 1000 \times 0.1156$

$= 115.6 \simeq 116$ (approx.).

Example 7. *The life of army shoes is normally distributed with mean 8 months and standard deviation 2 months. If 5000 pairs are, issued, how many pairs would be expected to need replacement after 12 months ?*

Solution. Here, $\mu = 8$, $\sigma = 2$, $N = 5000$

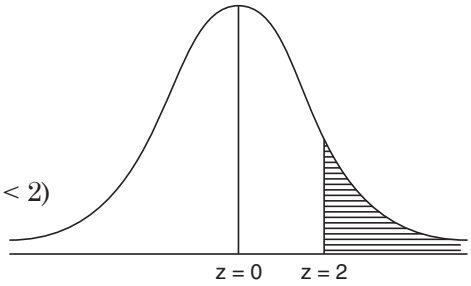
$$Z = \frac{x - \mu}{\sigma} = \frac{12 - 8}{2} = \frac{4}{2} = 2$$

$P(x > 12) = P(Z > 2)$

$= P(0 < Z < \infty) - P(0 < Z < 2)$

$= 0.5 - 0.4772$

$= 0.0228$



Number of pairs whose life is more than 12 months

$= 5000 \times 0.0228 = 114$

Replacement after 12 months = 5000 – 114 = 4886 Pairs.

Example 8. *Assuming that the diameters of 1000 brass plugs taken consecutively from a machine form a normal distribution with mean 0.7515 inches and standard deviation 0.0020 inches. How many of the plugs are likely to be rejected if the diameter is to be 0.752 ± 0.004 inches ?*

Solution. Here, $N = 1000$, $\mu = 0.7515$, $\sigma = 0.0020$

Least diameter of non-defective plug

$= 0.752 - 0.004 = 0.748$ inches

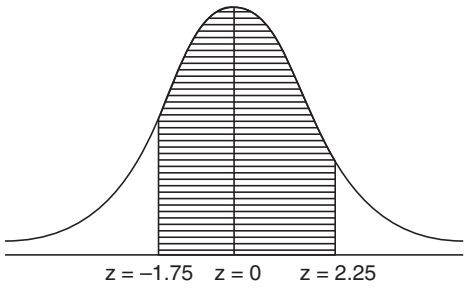
Greatest diameter of non-defective plug

$= 0.752 + 0.004$

$= 0.756$ inches

When $x = 0.748$, $Z = \frac{0.748 - 0.7515}{0.0020}$

$= -\frac{0.0035}{0.0020} = -1.75$



When $x = 0.756$, $Z = \frac{0.756 - 0.7515}{0.0020} = \frac{0.0045}{0.0020} = 2.25$

$P(0.748 \le x \le 0.756) = P(-1.75 \le Z \le 2.25)$

$= P(-1.75 \le Z \le 0) + P(0 \le Z \le 2.25)$

$= P(0 \le Z \le 1.75) + P(0 \le Z \le 2.25)$

$= 0.4599 + 0.4878 = 0.9477$

Number of plugs to be accepted = $1000 \times 0.9477 = 947.7 \simeq 948$ (approx.)

Number of plugs likely to be rejected = 1000 – 948 = 52.

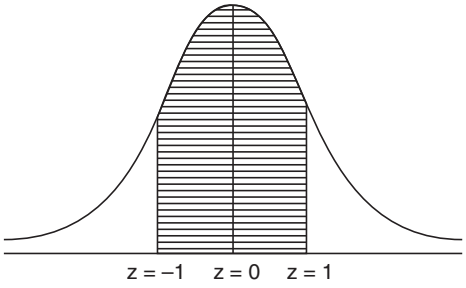
Example 9. *A manufacturer knows from experience that the resistance of resistors he produces is normal with mean 100 ohms and standard deviation 2 ohms. What percentage of resistors will have resistance between 98 ohms and 102 ohms ?*

Solution. Here, $\mu = 100, \sigma = 2$

$$Z = \frac{x - \mu}{\sigma} = \frac{x - 100}{2}$$

When $x = 98$, $Z = \frac{98 - 100}{2} = -\frac{2}{2} = -1$

When $x = 102$, $Z = \frac{102 - 100}{2} = \frac{2}{2} = 1$



$P(98 < x < 102) = P(-1 < Z < 1) = P(-1 \le Z < 0) + P(0 \le Z \le 1)$

$= P(0 \le Z \le 1) + P(0 \le Z \le 1) = 2P(0 \le Z \le 1)$

$= 2 \times 0.3413 = 0.6826$

Resistors having resistance between 98 ohms and 102 ohms = 68.26%.

Example 10. *In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.*

Sol. Let μ be the mean and σ be the standard deviation.

When $x = 45$, $Z_1 = \frac{45 - \mu}{\sigma}$

When $x = 64$, $Z_2 = \frac{64 - \mu}{\sigma}$

Area between 0 and $Z_1 = 0.50 - 0.31 = 0.19$

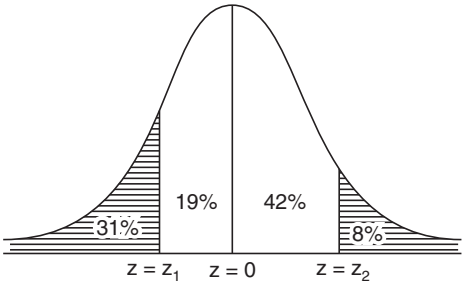
From the table, when area is 0.19,

$Z_1 = -0.496$ $(Z_1 < 0)$

Area between 0 and $Z_2 = 0.5 - 0.08 = 0.42$

From the table, when area is 0.42,

$Z_2 = 1.405$



We have

$$-0.496 = \frac{45 - \mu}{\sigma} \quad \text{and} \quad 1.405 = \frac{64 - \mu}{\sigma}$$

Now on solving for μ and σ, we have

$\mu = 50$ and $\sigma = 10$.

Example 11. *Fit a normal curve to the following data :*

Length of line (in cm)	*4*	*6*	*8*	*10*	*12*	*14*	*16*	*18*	*20*	*22*	*24*
Frequency	*1*	*7*	*15*	*22*	*35*	*43*	*38*	*20*	*13*	*5*	*1*

Solution. First find the mean μ and standard deviation σ as follows :

Let assumed mean $A = 14$

NOTES

Length of line (in cm) x	*Frequency* f	$d = x - A = x - 14$	fd	fd^2
4	1	– 10	– 10	100
6	7	– 8	– 56	448
8	15	– 6	– 90	540
10	22	– 4	– 88	352
12	35	– 2	– 70	140
14 (A)	43	0	0	0
16	38	2	76	152
18	20	4	80	320
20	13	6	78	468
22	5	8	40	320
24	1	10	10	100
Total	$\Sigma f = 200$		$\Sigma fd = -30$	$\Sigma fd^2 = 2940$

$$\text{Mean } (\mu) = A + \frac{\Sigma fd}{\Sigma f} = 14 + \frac{(-30)}{200} = 14 - 0.15 = 13.85$$

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2} = \sqrt{\frac{2940}{200} - \left(-\frac{30}{200}\right)^2}$$

$$= \sqrt{14.7 - 0.0225} = \sqrt{14.6775} = 3.83$$

Hence, the equation of the normal curve is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{(3.83)\sqrt{2\pi}} e^{-\frac{(x-13.85)^2}{29.355}}.$$

EXERCISE 2.3

1. On a final examination in Statistics, the mean was 72, and the standard deviation was 15. Determine the standard scores of students receiving grades :

(*i*) 60 (*ii*) 93 (*iii*) 72.

2. Find the area under the normal curve in each of the cases :

(*i*) Z = 0 and Z = 1.2 (*ii*) Z = – 0.68 and Z = 0

(*iii*) Z = – 0.46 and Z = 2.21 (*iv*) Z = 0.81 and Z = 1.94

(*v*) To the left of Z = – 0.6 (*vi*) Right of Z = – 1.28.

3. Find the value of Z in each of the cases

(*i*) area between 0 and Z is 0.3770 (*ii*) area to the left of Z is 0.8621

4. If X is a normal variate with mean 12 and standard deviation 4, then find

(*i*) $P(X \geq 20)$ (*ii*) $P(X \leq 20)$

5. The scores of candidates in a certain test are normally distributed with mean 500 and standard deviation 100. What percentage of candidates receives the scores between 350 and 550 ?

6. Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches square. How many soldiers in a regiment of 10,000 would you expect to be over 6 feet tall?

7. In a sample of 1000 items, the mean weight and standard deviation are 45 kgs and 15 kgs respectively. Assuming the distribution to be normal, find the number of items weighing between 40 kgs and 60 kgs.

8. A workshop produces 2000 units per day. The average weight of units is 130 kgs with a standard deviation of 10 kgs. Assuming normal distribution, how many units are expected to weight less than 142 kgs ?

9. The mean of a normal distribution is 50 and 5% of the values are greater than 60. Find the standard deviation of the distribution.

10. The time taken to complete a particular type of job is distributed approximately normal with a mean of 1.8 hours and a standard deviation 0.1 hour. If 'Normal time work' finishes at 6.00 p.m. and a job is started at 4.00 p.m. then, what is the probability that the job will need overtime payments ?

11. The marks of the students in a class are normally distributed with mean 70 and standard deviation 5. If the instructor decides to give 'A' grade to the top 15% students of the class, how many marks a student must get to be able to get 'A' grade ?

12. Find the values of mean and standard deviation from the following data relating to a normal distribution ?

10% of the items are under 40

95% of the items are under 75.

13. In a sample of 240 workers in a factory, the mean and standard deviation of wages were ₹ 113.50 and ₹ 30.30 respectively. Find the percentage of workers getting wages between ₹ 90 and ₹ 170 in the whole factory assuming that the wages are normally distributed.

14. In a distribution exactly normal, 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution ?

15. Fit a normal curve to the following data :

Variable (*r*)	0	1	2	3	4	5
Frequency (*f*)	10	14	19	8	5	4

Answers

1. (*i*) – 0.8 (*ii*) 1.4 (*iii*) 0

2. (*i*) 0.3849 (*ii*) 0.2518 (*iii*) 0.6637 (*iv*) 0.1828
(*v*) 0.2743 (*vi*) 0.8997 **3.** (*i*) $z = \pm 1.16$ (*ii*) Z = 1.09

4. (*i*) 0.0228 (*ii*) 0.9772 **5.** 62.47% **6.** 1251

7. 471 approx. **8.** 1770 **9.** 6.1 **10.** 0.0228

11. 75.18 **12.** 55.32, 11.97 **13.** 75% **14.** 50.3, 10.33

15. $f(x) = \frac{1}{1.40\sqrt{2\pi}} e^{-\frac{(x-1.93)^2}{3.92}}$.

NOTES

3. STATISTICAL DECISION THEORY

STRUCTURE

Introduction
Elements of a Decision Problem
Types of Decision Making Environment
Decision Making Under Uncertainty
Decision Making Under Risk
Decision Tree

3.1. INTRODUCTION

Decision analysis involves the use of a rational process for selecting the best of several alternatives. The goodness of a selected alternative depends on the quality of the data used in describing the decision situation. A decision making process falls into one of the three categories ; decision making under certainty, decision making under uncertainty and decision making under risk.

Nowadays, management students, businessman, engineers, persons from industries and government, etc. are giving much emphasis over decision making under conditions of uncertainty as mostly situations involves making choices under uncertainty. The study of making decisions to choose the best among a number of alternative courses of action is known as decision theory or statistical decision theory.

3.2. ELEMENTS OF A DECISION PROBLEM

There are certain essential elements which are common to all decision making categories :

1. **The decision maker.** The decision maker is charged with the responsibility for making the decisions. The decision maker can be an individual, a group of individuals, any company, an industrial body, etc.

2. **Acts.** The acts are the alternative courses of action or strategies that are available to decision maker.

3. **Events.** The events also known as state of nature. The events identify the occurrences which are not under the control of decision maker and which determines the level of success for a given act.

4. **Pay-off.** Each combination of a course of action and an event is associated with a pay-off. It is a quantitative measure of the value to the decision maker of the outcomes. It measures the net benefit to the decision maker from a given combination of course of action and an event. The pay-off usually represents the net monetary gain (profit), but some other measures can also be used, as cost is negative profit.

5. **Pay-off table.** Suppose the problem under consideration has m possible events (state of nature) denoted by $E_1, E_2, ..., E_m$ and n alternative acts (strategies) denoted by $A_1, A_2, ..., A_n$. Then the pay-off corresponding to strategy A_j of the decision maker under the state of nature E_i will be denoted by p_{ij} ($i = 1, 2, ..., m$; $j = 1, 2, ..., n$).

The totality of mn pay-offs arranged in a tabular form is known as pay-off table.

Events (States of nature)	*Pay-off (₹) courses of actions/acts/strategies*			
	A_1	A_2	.....	A_n
E_1	p_{11}	p_{12}	.....	p_{1n}
E_2	p_{21}	p_{22}	.....	p_{2n}
⋮	⋮	⋮		⋮
E_m	p_{m1}	p_{m2}	.....	p_{mn}

6. **Regret or opportunity loss table.** An opportunity loss has been defined to be the difference between the highest possible profit for a state of nature and the actual profit obtained for a particular action taken, *i.e.* an opportunity loss is the loss incurred due to failure of not adopting the best possible action. For a given state of nature, the opportunity loss of a course of action is the difference between the pay-off of that course of action and the pay-off for the best course of action that could have been selected.

Events (States of nature)	*Opportunity loss (₹) courses of actions/acts/strategies*			
	A_1	A_2	.....	A_n
E_1	$M_1 - p_{11}$	$M_1 - p_{12}$	.....	$M_1 - p_{1n}$
E_2	$M_2 - p_{21}$	$M_2 - p_{22}$	.....	$M_2 - p_{2n}$
⋮	⋮	⋮		⋮
E_m	$M_m - p_{m1}$	$M_m - p_{m2}$	.....	$M_m - p_{mn}$

where $M_1, M_2, ..., M_m$ are the maximum of these quantities respectively.

3.3. TYPES OF DECISION MAKING ENVIRONMENT

NOTES

There are three categories of decision making environment.

1. **Decision making under certainty.** In this environment the decision maker knows with certainty the consequence of every alternative or decision choice. There will be only one outcome for each alternative. Examples of such decision problems are linear programming, dynamic programming, transportation, assignment, integer programming, etc.

2. **Decision making under uncertainty.** In this environment, the decision maker cannot assess the outcome probability with confidence. In otherwords, if the information about the outcomes is incomplete and the available information cannot be described in terms of probability density. Decisions under uncertainty refer to situations where more than one outcome can result from any single decision.

3. **Decision making under risk.** In this environment the pay-offs associated with each decision alternative are usually described by probability distributions. For this reason, decision making under risk is usually based on the expected monetary value criterion or expected opportunity loss of the expected pay-off.

3.4. DECISION MAKING UNDER UNCERTAINTY

Different rules for making a decision under such environment are as follows :

1. **Maximax or Minimin Criterion (Criterion of Optimism).** This criterion is based upon 'extreme optimism'. The basic steps of this criterion are as follows :

(*i*) Determine the maximum possible pay-off for each alternative.

(*ii*) Choose that alternative which corresponds to the maximum of the above maximum pay-offs.

In decision problems dealing with costs, the minimum for each alternative is determined and then the alternative which minimizes the above minimum cost is selected.

2. **Maximin or Minimax Criterion (Criterion of Pessimism).** This criterion is based upon the 'conservative approach' to assume that the worst possible is going to happen. The basic steps of this criterion are as follows :

(*i*) Determine the minimum possible pay-off for each alternative.

(*ii*) Choose that alternative which corresponds to the maximum of the above minimum pay-offs.

In decision problems dealing with costs, the maximum for each alternative is determined and then the alternative which minimizes the above maximum cost is selected.

3. **Laplace Criterion (Equally Likely Decisions).** The Laplace criterion uses all the information by assigning equal probabilities to all events of each alternative, as there is no information about probabilities of occurrence. The basic steps of this criterion are as follows :

(*i*) Assign equal probabilities $\left(\frac{1}{n}\right)$ to each pay-off of a strategy (having n pay-offs).

(*ii*) Determine the expected pay-off value for each alternative by multiplying each pay-off by its probability and then adding.

(*iii*) Choose that alternative which corresponds to the maximum of the expected pay-offs.

In decision problems dealing with costs, select that alternative which corresponds to the minimum of the expected pay-offs.

4. **Savage Criterion (Criterion of Regret).** The savage criterion is based on the concept of regret (or opportunity loss). This criterion also known as minimax regret criterion. The basic steps of this criterion are as follows :

(*i*) Construct the regret table.

regret (opportunity loss)

$$= \begin{cases} \text{pay-off} - \max.\ \text{pay-off}\ ; \text{if the pay-offs represent profits} \\ \text{pay-off} - \min.\ \text{pay-off}\ ; \text{if the pay-offs represent costs} \end{cases}$$

(*ii*) Determine the maximum regret for each alternative.

(*iii*) Choose the alternative with minimum regret out of these maximum regrets.

5. **Hurwicz Criterion.** The Hurwicz criterion is based on the concept that the decision makers are neither completely pessimistic nor completely optimistic, but are a combination of the two extremes. Therefore, we should give attention to both. The basic steps of this criterion are as follows :

(*i*) Choose an appropriate degree of optimism (or pessimism) of the decision maker. Let α ($0 \le \alpha \le 1$) be his degree of optimism (so $1 - \alpha$ is his degree of pessimism).

(*ii*) Determine the maximum as well as minimum pay-off for each alternative and obtain the quantities D (decision index) as

$D = \alpha$.maximum pay-off + $(1 - \alpha)$. minimum pay-off

for each alternative.

(*iii*)Choose the maximum value of D when profits are given (Choose the minimum value of D when costs are given).

SOLVED EXAMPLES

Example 1. *Given the following profit pay-off table :*

Strategy	*States of nature* Pay-off (in ₹)			
	S1	*S2*	*S3*	*S4*
A1	*16*	*10*	*12*	*7*
A2	*13*	*12*	*9*	*9*
A3	*11*	*14*	*15*	*14*

Which strategy should the decision maker choose on the basis of

(i) Maximin criterion *(ii) Maximax criterion*

(iii) Minimax regret criterion *(iv) Laplace criterion ?*

NOTES

Solution.

(*i*) **Maximin criterion :**

Strategy	*States of nature*				*Minimum for each strategy*
	S1	*S2*	*S3*	*S4*	
A1	16	10	12	7	7
A2	13	12	9	9	9
A3	11	14	15	14	11 (Max.)

Using maximin criterion, maximum of these minimum is 11 corresponding to strategy A3. So A3 should be selected.

(*ii*) **Maximax criterion :**

Strategy	*States of nature*				*Minimum for each strategy*
	S1	*S2*	*S3*	*S4*	
A1	16	10	12	7	16 (Max)
A2	13	12	9	9	13
A3	11	14	15	14	15

Using maximax criterion, maximum of these maximum is 16 corresponding to strategy A1. So A1 should be selected.

(*iii*) **Minimax regret criterion :**

Regret table

Strategy	*States of nature*				*Maximum Regret*
	S1	*S2*	*S3*	*S4*	
A1	16–16 = 0	14–10 = 4	15–12 = 3	14–7 = 7	7
A2	16–13 = 3	14–12 = 2	15–9 = 6	14–9 = 5	6
A3	16–11 = 5	14–14 = 0	15–15 = 0	14–14 = 0	5 (Min.)

Using minimax regret criterion, minimum of these maximum regrets is 5 corresponding to strategy A3. So A3 should be selected.

(*iv*) **Laplace criterion :**

Here, $p = 1/4$

$$\text{EMV (Strategy A1)} = \frac{1}{4} \times 16 + \frac{1}{4} \times 10 + \frac{1}{4} \times 12 + \frac{1}{4} \times 7$$

$$= \frac{1}{4}(16 + 10 + 12 + 7) = \frac{45}{4} = 11.25$$

$$\text{EMV (Strategy A2)} = \frac{1}{4}(13 + 12 + 9 + 9) = \frac{43}{4} = 10.75$$

$$\text{EMV (Strategy A3)} = \frac{1}{4}(11 + 14 + 15 + 14) = \frac{54}{4} = 13.5$$

Since the EMV is maximum for strategy A3. So A3 should be selected.

Example 2. *The research department of Hindustan Lever has recommended to the marketing department to launch a shampoo of three different types. The marketing manager has to decide one of the types of shampoo to be launched under the following estimated pay-offs various levels of sales :*

Types of shampoo	*Estimated levels of sales (units)*		
	15,000	*10,000*	*5,000*
Egg shampoo	*30*	*10*	*10*
Clinic shampoo	*40*	*15*	*5*
Delux shampoo	*55*	*20*	*3*

What should be the marketing manager decision using

(*i*) *Maximin criterion* (*ii*) *Minimax criterion*

(*iii*) *Maximax criterion* (*iv*) *Laplace criterion*

(*v*) *Minimax regret criterion ?*

Solution. (*i*) **Maximin criterion :**

Types of shampoo	*Estimated levels of sales (unit)*			*Minimum of each decision*
	15,000	*10,000*	*5,000*	
Egg	30	10	10	10 (Max.)
Clinic	40	15	5	5
Delux	55	20	3	3

Using maximin criterion, maximum of these minimum is 10 corresponding to Egg shampoo. So Egg shampoo should be launched.

(*ii*) **Minimax criterion :**

Types of shampoo	*Estimated levels of sale (units)*			*Maximum of each decision*
	15,000	*10,000*	*5,000*	
Egg	30	10	10	30 (Min)
Clinic	40	15	5	40
Delux	55	20	3	55

Using minimax criterion, minimum of these maximum is 30 corresponding to Egg shampoo. So Egg shampoo should be launched.

(*iii*) **Maximax criterion :**

Types of shampoo	*Estimated levels of sale (units)*			*Maximum of each decision*
	15,000	*10,000*	*5,000*	
Egg	30	10	10	30
Clinic	40	15	5	40
Delux	55	20	3	55 (Max.)

Using maximax criterion, maximum of these maximum is 55 corresponding to Delux shampoo. So Delux shampoo should be launched.

NOTES

(*iv*) **Laplace criterion :** Here, $p = \frac{1}{3}$

$$\text{EMV (Egg shampoo)} = \frac{1}{3} \times 30 + \frac{1}{3} \times 10 + \frac{1}{3} \times 10$$

$$= \frac{1}{3}(30 + 10 + 10) = \frac{50}{3} = 16.67$$

$$\text{EMV (Clinic shampoo)} = \frac{1}{3}(40 + 15 + 5) = \frac{60}{3} = 20$$

$$\text{EMV (Delux shampoo)} = \frac{1}{3}(55 + 20 + 3) = \frac{78}{3} = 26$$

Since the EMV is maximum for Delux shampoo. So Delux shampoo should be launched.

(*v*) **Minimax regret criterion :**

Regret table

Types of shampoo	*Estimated levels of sale (units)*			*Maximum regret*
	15,000	*10,000*	*5,000*	
Egg	55–30 = 25	20–10 = 10	10–10 = 0	25
Clinic	55–40 = 15	20–15 = 5	10–5 = 5	15
Delux	55–55 = 0	20–20 = 0	10–3 = 7	7 (Min.)

Using minimax regret criterion, minimum of these maximum regrets is 7 corresponding to Delux shampoo. So Delux shampoo should be launched.

Example 3. *A farmer wants to plan which of the three crops he should plant on his 100 acre farm. The profit of each crop depends upon the rainfall during the growing season. The rainfall could be high, medium and low. The estimated profit of the farmer for each of the crops is shown in the table :*

Rainfall	*Estimated Conditional Profit*		
	Crop A	*Crop B*	*Crop C*
High	*6000*	*3000*	*7000*
Medium	*4000*	*4500*	*4000*
Low	*2000*	*5000*	*5000*

The farmer decides to plant only one crop, which would be his best crop using the following :

(*i*) *Maximax criterion* (*ii*) *Maximin criterion*

(*iii*) *Laplace criterion* (*iv*) *Minimax regret criterion.*

Solution. (*i*) **Maximax criterion :**

Type of crop	*Estimated Conditional Profit*			*Maximum of each crop*
	Rainfall			
	High	*Medium*	*Low*	
Crop A	6000	4000	2000	6000
Crop B	3000	4500	5000	5000
Crop C	7000	4000	5000	7000 (Max.)

Using maximax criterion, maximum of these maximum is 7000 corresponding to crop C.

So crop C is the best crop.

(*ii*) **Maximin criterion :**

Type of crop	*Estimated Conditional Profit*			*Minimum of each crop*
	Rainfall			
	High	*Medium*	*Low*	
Crop A	6000	4000	2000	2000
Crop B	3000	4500	5000	3000
Crop C	7000	4000	5000	4000 (Max.)

Using maximin criterion, maximum of these minimum is 4000 corresponding to crop C.

So crop C is the best crop.

(*iii*) **Laplace criterion :** Here, $p = \frac{1}{3}$

$$\text{EMV (Crop A)} = \frac{1}{3} \times 6000 + \frac{1}{3} \times 4000 + \frac{1}{3} \times 2000$$

$$= \frac{1}{3}(6000 + 4000 + 2000) = \frac{12000}{3} = 4000$$

$$\text{EMV (Crop B)} = \frac{1}{3}(3000 + 4500 + 5000) = \frac{12500}{3} = 4166.67$$

$$\text{EMV (Crop C)} = \frac{1}{3}(7000 + 4000 + 5000) = \frac{16000}{3} = 5333.33$$

Since the EMV is maximum for crop C. So crop C is the best crop.

(*iv*) **Minimax regret criterion :**

Regret table

Types of crop	*Rainfall*			*Maximum regret*
	High	*Medium*	*Low*	
Crop A	7000–6000 = 1000	4500–4000 = 500	5000–2000 = 3000	3000
Crop B	7000–3000 = 4000	4500–4500 = 0	5000–5000 = 0	4000
Crop C	7000–7000 = 0	4500–4000 = 500	5000–5000 = 0	500 (Min.)

Using minimax regret criterion, minimum of these maximum regrets is 500 corresponding to crop C. So crop C is best crop.

NOTES

Example 4. *Consider the following pay-off (profit) matrix :*

Alternative	*Events*			
	E1	*E2*	*E3*	*E4*
A1	*5*	*10*	*18*	*25*
A2	*8*	*7*	*8*	*23*
A3	*21*	*18*	*12*	*21*
A4	*30*	*22*	*19*	*15*

Find optimum alternative using Hurwicz criterion with $\alpha = 0.75$.

Solution. Here, $\alpha = 0.75$ so $(1 - \alpha) = 1 - 0.75 = 0.25$

Alternative	*Maximum pay-off (i)*	*Minimum pay-off (ii)*	$D = \alpha \times (i) + (1 - \alpha)\ (ii)$
A1	25	5	$25 \times 0.75 + 5 \times 0.25 = 20$
A2	23	7	$23 \times 0.75 + 7 \times 0.25 = 19$
A3	21	12	$21 \times 0.75 + 12 \times 0.25 = 18.75$
A4	30	15	$30 \times 0.75 + 15 \times 0.25 = 26.25$

According to Hurwicz criterion, maximum value of D is 26.25 corresponding to A4. So A4 is optimum alternative.

3.5. DECISION MAKING UNDER RISK

Different criterion for making a decision under such environment are as follows:

1. **Expected Monetary Value (EMV) Criterion.** The expected monetary value criterion seeks the maximization of expected profit or the minimization of expected cost. The basic steps of this criterion are as follows :

(*i*) Construct the pay-off table listing all possible courses of actions and events (states of nature), along with the corresponding event probabilities.

(*ii*)Determine the expected conditional profit values for each course of action.

(*iii*) Determine EMV for each course of action (strategy) by

$$\text{EMV}\ (A_i) = p_{i1}\ P(E_1) + p_{i2}\ P(E_2) + \ldots + p_{im}\ P(E_m)$$

(*iv*) Choose that course of action (strategy) having highest EMV.

2. **Expected Opportunity Loss (EOL) Criterion.** An alternative approach to maximizing expected monetary value (EMV) is to minimize expected opportunity loss (EOL). The basic steps of this criterion are as follows :

(*i*) Construct the opportunity loss table listing all possible courses of actions and events (states of nature), along with the corresponding event probabilities.

(*ii*) Determine the conditional opportunity loss values for each event.

(*iii*) Determine the expected conditional opportunity loss values and sum these values to get the expected opportunity loss (EOL) for each course of action by

$$\text{EOL}\ (A_j) = (M_1 - p_{1j})\ P(E_1) + (M_2 - p_{2j})\ P(E_2) + \ldots + (M_m - p_{mj})\ P\ (E_m)$$

$$(j = 1, 2, \ldots, n)$$

(*iv*) Choose that course of action (strategy) having lowest EOL.

3. **Expected Value of Perfect Information (EVPI).** The expected value with perfect information is the expected or average return, in the long run, if we have perfect information before a decision is made. The EVPI may be defined as the maximum amount spend by the decision maker to get perfect (additional) information. Expected pay-off under perfect information (EPPI) can be calculated by finding the sum of product of pay-off of best outcome of each state of nature and its probability of occurrence.

The expected value of perfect information (EVPI) is the expected outcome with perfect information minus the expected outcome without perfect information (maximum EMV).

$$EVPI = EPPI - \text{max. EMV}$$

SOLVED EXAMPLES

Example 1. *A management is faced with the problem of choosing one of three products for manufacturing. The potential demand for each product may turn out to be good, moderate or poor. The probabilities for each of the states of nature were estimated as follows :*

Product	*Nature of demand*		
	Good	*Moderate*	*Poor*
X	*0.70*	*0.20*	*0.10*
Y	*0.50*	*0.30*	*0.20*
Z	*0.40*	*0.50*	*0.10*

The estimated profit or loss (in ₹) under the three states may be taken as :

Product	*Good*	*Moderate*	*Poor*
X	*300,000*	*200,000*	*100,000*
Y	*600,000*	*300,000*	*200,000*
Z	*400,000*	*100,000*	*– 150,000 (loss)*

Prepare the expected value table and advice the management about the choice of the product.

Solution.

Nature of demand	*Expected pay-off (in ₹ Lacs) for various acts*								
	X			*Y*			*Z*		
	x_{1j}	p_{1j}	$x_{1j}\,p_{1j}$	x_{2j}	p_{2j}	$x_{2j}\,p_{2j}$	x_{3j}	p_{3j}	$x_{3j}\,p_{3j}$
Good	3	0.7	2.1	6	0.5	3.0	4	0.4	1.6
Moderate	2	0.2	0.4	3	0.3	0.9	1	0.5	0.5
Poor	1	0.1	0.1	2	0.2	0.4	– 1.5	0.1	– 0.15
EMV = $\Sigma x_{ij}\,p_{ij}$			2.6			4.3			1.95

Since the EMV is maximum for product Y, so Y should be selected as the best product.

NOTES

Example 2. *Pay-offs of three acts X, Y, Z and the states of nature P, Q, R are as follows :*

State of nature	*Pay-offs (₹) (Acts)*		
	X	*Y*	*Z*
P	*– 120*	*– 80*	*100*
Q	*200*	*400*	*– 300*
R	*260*	*– 260*	*600*

The probabilities of the states of nature are 0.3, 0.5 and 0.2 respectively. Tabulate the expected monetary values (EMVs) for the above data and state which can be selected as the best act.

Solution.

State of nature	*Pay-offs (₹) (Acts)*								
	X			*Y*			*Z*		
	x_{1j}	p_{1j}	$x_{1j}p_{1j}$	x_{2j}	p_{2j}	$x_{2j}p_{2j}$	x_{3j}	p_{3j}	$x_{3j}p_{3j}$
P	– 120	0.3	– 36	– 8.0	0.3	– 24	100	0.3	30
Q	200	0.5	100	400	0.5	200	– 300	0.5	– 150
R	260	0.2	52	– 260	0.2	– 52	600	0.2	120
EMV = $\Sigma x_{ij}p_{ij}$			116			124			0

Since the EMV is maximum for act Y, so Y should be selected as the best act.

Example 3. *A newspaper distributor assigns probabilities to the demand for a magazine as follows :*

Copies demanded	*1*	*2*	*3*	*4*
Probability	*0.4*	*0.3*	*0.2*	*0.1*

A copy of magazine sells for ₹ 7 and costs ₹ 6. What can be the maximum possible expected monetary value (EMV) if the distributor can return unsold copies for ₹ 5 each ?

Solution. Cost of a magazine = ₹ 6

Selling price of a magazine = ₹ 7

Profit per magazine = ₹ (7 – 6) = ₹ 1

Loss on each unsold magazine

= ₹ (6 – 5) = ₹ 1

$$\text{Conditional profit} = \begin{cases} 1 \times S = S & \text{if } D \geq S \\ 1 \times D - 1 \times (S - D) = 2D - S & \text{if } D < S \end{cases},$$

where D = no. of magazines demanded

S = no. of magazines in stock

The resulting pay-off and corresponding expected pay-offs are as follows :

Event (Demand)	*Probability*	*Conditional pay-off (₹) Act (Stock)*				*Expected pay-off (₹) Act (Stock)*			
		1	*2*	*3*	*4*	*1*	*2*	*3*	*4*
D	*(i)*	*(ii)*	*(iii)*	*(iv)*	*(v)*	*(i) × (ii)*	*(i) × (iii)*	*(i) × (iv)*	*(i) × (v)*
1	0.4	1	0	– 1	– 2	0.4	0	– 0.4	– 0.8
2	0.3	1	2	1	0	0.3	0.6	0.3	0
3	0.2	1	2	3	2	0.2	0.4	0.6	0.4
4	0.1	1	2	3	4	0.1	0.2	0.3	0.4
					EMV	1.0	1.2	0.8	0

Since the EMV is maximum for act (stock) 2, so the optimum act for the distributor would be to stock 2 copies of magazine.

Example 4. *The following pay-off table is given :*

Acts	*Events*			
	E1	*E2*	*E3*	*E4*
A1	*40*	*200*	*– 200*	*100*
A2	*200*	*0*	*200*	*0*
A3	*0*	*100*	*0*	*150*
A4	*– 50*	*400*	*100*	*0*

Suppose that the probabilities of the events are :

$P(E_1) = 0.20$, $P(E_2) = 0.15$, $P(E_3) = 0.40$ and $P(E_4) = 0.25$. Calculate the expected pay-off and the expected loss of each action. Find the optimum act using EMV and EOL criterion.

Solution. Computation of expected pay-offs

Event (Demand)	*Probability*	*Conditional pay-off (₹) Act (Stock)*				*Expected pay-off (₹) Act (Stock)*			
		1	*2*	*3*	*4*	*1*	*2*	*3*	*4*
D	*(i)*	*(ii)*	*(iii)*	*(iv)*	*(v)*	*(i) × (ii)*	*(i) × (iii)*	*(i) × (iv)*	*(i) × (v)*
E1	0.20	40	200	0	– 50	8	40	0	– 10
E2	0.15	200	0	100	400	30	0	15	60
E3	0.40	– 200	200	0	100	– 80	80	0	40
E4	0.25	100	0	150	0	25	0	37.5	0
					EMV	– 17	120	52.5	90

Since the EMV is maximum for act A2, so A2 is the optimum act.

NOTES

Computation of expected loss

Event	*Probability*	*Opportunity loss*				*Expected loss*			
		Act				*Act*			
		A1	*A2*	*A3*	*A4*	*A1*	*A2*	*A3*	*A4*
	(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*	*(i) × (ii)*	*(i) × (iii)*	*(i) × (iv)*	*(i) × (v)*
E1	0.20	200 – 40 = 160	200 – 200 = 0	200 – 0 = 200	200 + 50 = 250	32	0	40	50
E2	0.15	400–200 = 200	400–0 = 400	400–100 = 300	400–400 = 0	30	60	45	0
E3	0.40	200 + 200 = 400	200–200 = 0	200–0 = 200	200–100 = 100	160	0	80	40
E4	0.25	150–100 = 50	150–0 = 150	150–150 = 0	150–0 = 150	12.5	37.5	0	37.5
					EOL	234.5	97.5	165	127.5

Since the EOL is minimum for act A2, so A2 is the optimum act.

Example 5. *A grocery with a bakery department is faced with the problem of how many cakes to buy in order to meet the day's demand. The grocer prefers not to sell day-old goods in competition with fresh products ; leftover cakes are, therefore, a complete loss. On the other hand, if a customer desires a cake and all of them have been sold, the disappointed customer will buy elsewhere and the sales will be lost. The grocer has, therefore, collected information on the past sale on a selected 100 day period as follows:*

Sales per day	*No. of days*	*Probability*
25	*10*	*0.10*
26	*30*	*0.30*
27	*50*	*0.50*
28	*10*	*0.10*

A cake costs ₹ 80 and sells for ₹ 100. Construct the pay-off table and the opportunity loss table. What is the optimum number of cakes that should be bought each day ?

Solution. Let A_i = alternative strategy (act) of stocking i cakes

E_j = a daily demand of j cakes state of nature (event)

Here, cost of a cake = ₹ 80

Selling price of a cake = ₹ 100

Profit per cake sold = ₹ (100 – 80) = ₹ 20

Loss on each unsold cake = ₹ 80

$$\text{Conditional pay-off} = \begin{cases} 20\,S & \text{if } D \geq S \\ 20\,D - 80\,(S - D) & \text{if } D < S, \\ = 100\,D - 80\,S & \end{cases}$$

where D = no. of cakes demanded

S = no. of cakes in stock

The resulting pay-off (conditional profit) are as follows :

Event (Demand) D	*Probability*	*Conditional pay-off (₹) Act (Stock)*			
		A1 : 25	*A2 : 26*	*A3 : 27*	*A4 : 28*
E1 : 25	0.10	500	420	340	260
E2 : 26	0.30	500	520	440	360
E3 : 27	0.50	500	520	540	460
E4 : 28	0.10	500	520	540	560

The expected conditional pay-off are computed as follows :

Event	*Probability*	*Expected conditional pay-off (₹)*			
		A1 : 25	*A2 : 26*	*A3 : 27*	*A4 : 28*
E1 : 25	0.10	50	42	34	26
E2 : 26	0.30	150	156	132	108
E3 : 27	0.50	250	260	270	230
E4 : 28	0.10	50	52	54	56
	EMV	500	510	490	420

Since the EMV is maximum for act A2, so 26 cakes should be bought (stocked) each day.

The conditional opportunity loss are computed as follows :

Event (Demand)	*Probability*	*Conditional opportunity loss (₹) Act (Stock)*			
		A1 : 25	*A2 : 26*	*A3 : 27*	*A4 : 28*
E1 : 25	0.10	0	80	160	240
E2 : 26	0.30	20	0	80	160
E3 : 27	0.50	40	20	0	80
E4 : 28	0.10	60	40	20	0

The expected conditional opportunity loss are computed as follows :

Event	*Probability*	*Expected conditional opportunity loss (₹)*			
		A1 : 25	*A2 : 26*	*A3 : 27*	*A4 : 28*
E1 : 25	0.10	0	8	16	24
E2 : 26	0.30	6	0	24	48
E3 : 27	0.50	20	10	0	40
E4 : 28	0.10	6	4	2	0
	EOL	32	22	42	112

Since the EOL is minimum for act A2, so 26 cakes should be bought (stocked) each day.

NOTES

Example 6. *A wholesaler of sports goods has an opportunity to by 5000 pairs of gloves that have been declared surplus by the government. The wholesaler will pay ₹ 50 per pair and can obtain ₹ 100 a pair by selling gloves to retailers. The price is well established, but the wholesaler is in doubt as to just how many pairs he will able to sell. Any gloves leftover, he can sell to discount outlets at ₹ 20 a pair. After a careful consideration of the past data, the wholesaler assigns probabilities to the demand as follows :*

Retailer's demand	*Probability*
1000 pairs	*0.6*
3000 pairs	*0.3*
5000 pairs	*0.1*

(*i*) *Compute the conditional monetary and expected monetary values.*

(*ii*) *Compute the expected profit with a perfect predicting device.*

(*iii*) *Compute the EVPI.*

Solution. Cost per pair = ₹ 50

Selling price per pair = ₹ 100

Profit per pair = ₹ 50 (on sold pair)

Disposal selling price = ₹ 20 (on unsold pair)

Loss on each unsold pair = ₹ (50 – 20) = ₹ 30

$$\text{Conditional pay-off (profit)} = \begin{cases} 50\,S & \text{if } D \geq S \\ 50\,D - 30\,(S - D) & \text{if } D < S \\ = 80\,D - 30\,S & \end{cases},$$

where D = no. of pairs demanded

S = no. of pairs stocked

(*i*) The resulting conditional pay-offs and corresponding expected pay-offs are computed as follows :

Retailer's demand D	*Probability*	*Conditional pay-offs (₹' 000) Stock per week*			*Expected pay-offs (₹ '000) Stock perweek*		
		1000 pairs	*3000 pairs*	*5000 pairs*	*1000 pairs*	*3000 pairs*	*5000 pairs*
1000 pairs	0.6	50	– 10	– 70	30	– 6	– 42
3000 pairs	0.3	50	150	90	15	45	27
5000 pairs	0.1	50	150	250	5	15	25
				EMV	50	54	10

(*ii*) The expected profit under perfect information (EPPI) is computed as follows:

Retailer's demand D	*Probability*	*Conditional pay-offs (₹'000) Stock per week*			*Under perfect information (₹ '000)*	
		1000 pairs	*3000 pairs*	*5000 pairs*	*Maximum pay-off (v)*	*Expected pay-off*
	(i)	*(ii)*	*(iii)*	*(iv)*	*[from (ii), (iii) and (iv)]*	*(i) × (v)*
1000 pairs	0.6	50	– 10	– 70	50	30
3000 pairs	0.3	50	150	90	150	45
5000 pairs	0.1	50	150	250	250	25
						EPPI = 100

(*iii*) EVPI = EPPI – max EMV = 100 – 54 = 46

Thus, EVPI = ₹ 46,000.

Example 7. *Pay-off of three acts, A1, A2 and A3 and state of nature X, Y, Z are as follows :*

State of nature	*Pay-off (₹) Acts*		
	A1	*A2*	*A3*
X	*– 20*	*– 50*	*200*
Y	*200*	*– 100*	*– 50*
Z	*400*	*600*	*300*

The probabilities of the state of nature are 0.3, 0.4 and 0.3 respectively. Calculate the EMV for the given data and select the best act. Also find the expected value of perfect information (EVPI).

Solution. Computation of expected pay-off

State of nature	*Probability*	*Pay-off (₹) Acts*			*Expected pay-off (₹) Acts*		
		A1	*A2*	*A3*	*A1*	*A2*	*A3*
X	0.3	– 20	– 50	200	– 6	– 15	60
Y	0.4	200	– 100	– 50	80	– 40	– 20
Z	0.3	400	600	300	120	180	90
				EMV	194	125	130

Since the EMV is maximum for act A1, so A1 is the best act.

NOTES

The expected profit under perfect information (EPPI) is computed as follows :

State of nature	*Probability*	*Pay-off (₹)* *Acts*			*Under perfect information (₹)*	
		A1	*A2*	*A3*	*Maximum pay-off*	*Expected pay-off*
X	0.3	– 20	– 50	200	200	200 × 0.3 = 60
Y	0.4	200	– 100	– 50	200	200 × 0.4 = 80
Z	0.3	400	600	300	600	600 × 0.3 = 180
						EPPI = 320

EVPI = EPPI – max EMV = 320 – 194 = 126

Thus, EVPI = ₹ 126.

3.6. DECISION TREE

A decision tree is a graphical representation of the various alternatives and sequence of events in a decision problem. Decision tree is a simple method for making a decision, where all the options are clearly open to the decision maker in concise form. Decision tree is beneficial for simple as well as complex decision making situations. Basically decision tree is drawn for those problems where more than one decisions are to be taken and the decision taken to one stage affects the subsequent decision.

There are some basic rules for drawing a decision tree.

(*i*) Identify all decisions to be made and the order in which they must be made.

(*ii*) Identify the chance events that can occur after each decision.

(*iii*) Develop a tree diagram showing the sequence of decisions and chance events. The tree is constructed starting from left and moving towards right. The 'square box' denotes a decision point at which the available strategies are considered. The 'circle' represents the chance node or event, the various state of nature or outcomes emanate from this.

(*iv*) Obtain probability estimate of the chances of each outcome's occurrence.

(*v*) Obtain estimates of the consequences of all possible outcomes or actions.

(*vi*) Calculate the expected value of all possible actions.

(*vii*) Select the action offering the most attractive expected value.

SOLVED EXAMPLES

Example 1. *XYZ Ltd. has invented a picture cell phone. It is faced with selecting one alternative out of the following strategies :*

(*i*) *Manufacture the cell phone*

(*ii*) *Take royalty from another manufacturer*

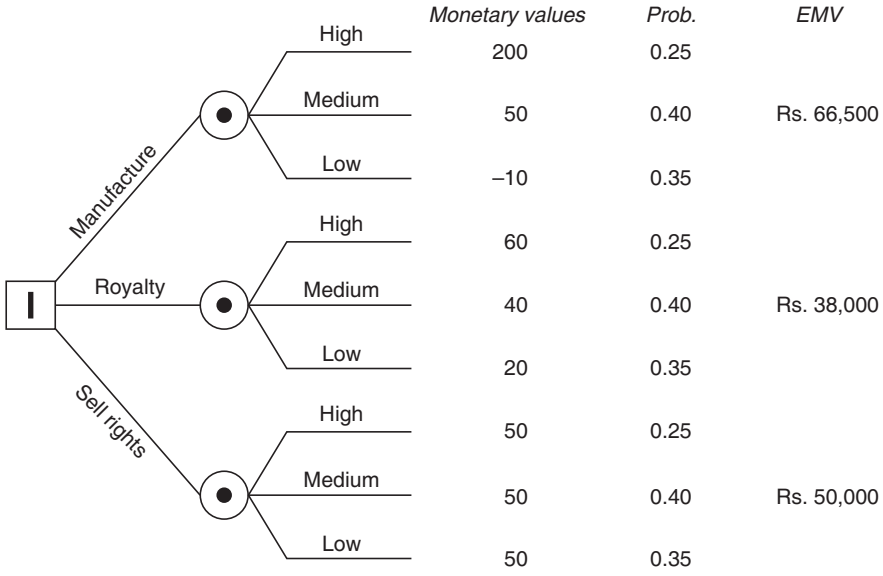
(*iii*) *Sell the rights for the invention and take a lumpsum amount.*

Profit in thousands of rupees which can be incurred and the probability associated with such alternative are shown in the followng table:

Event	*Probability*	*Manufacture*	*Royalty*	*Sell rights*
High	*0.25*	*200*	*60*	*50*
Medium	*0.40*	*50*	*40*	*50*
Low	*0.35*	*– 10*	*20*	*50*

Represent the company's problem in the form of the decision tree and suggest what decision the company should take to maximize profit.

Solution.



Thus, EMV for strategy (*i*) manufacture the cell phone is maximum. So the best decision by XYZ Ltd. is to manufacture the picture cell phone itself to get profit of ₹ 66,500.

Example 2. *A company is evaluating four alternative single-period investment opportunities whose returns are based on the state of the economy. The possible states of the economy and the associated probability distribution are as follows :*

State	*Fair*	*Good*	*Great*
Probability	*0.2*	*0.5*	*0.3*

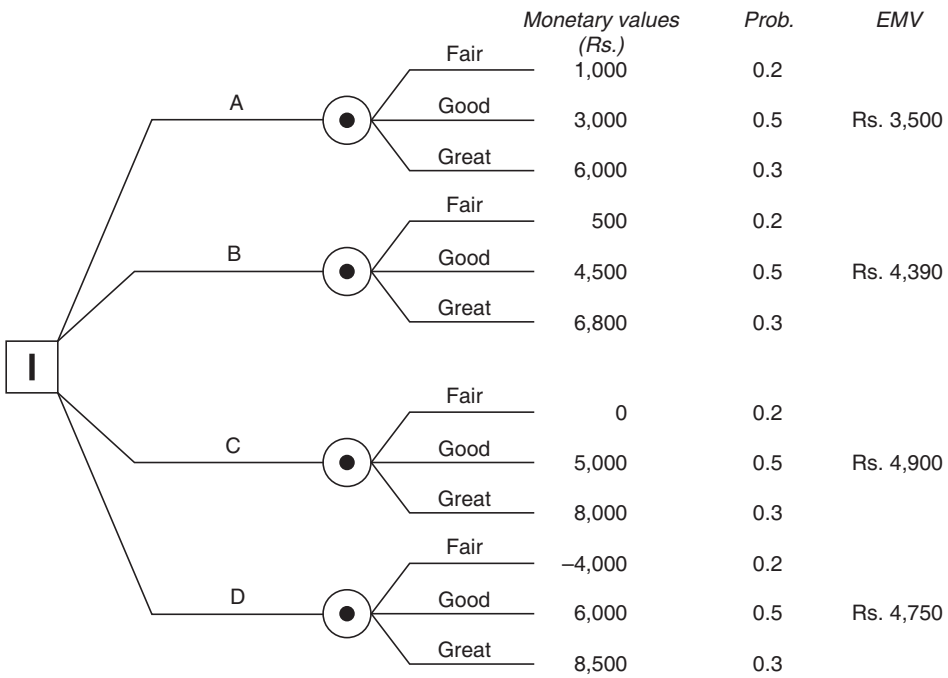
The returns for each investment opportunity and each state of the economy are as follows :

Alternative	*State of Economy*		
	Fair (₹)	*Good* (₹)	*Great* (₹)
A	*1,000*	*3,000*	*6,000*
B	*500*	*4,500*	*6,800*
C	*0*	*5,000*	*8,000*
D	*– 4,000*	*6,000*	*8,500*

NOTES

Using the decision tree approach, determine the expected return for each alternative. Which alternative investment proposal would you recommend if the expected monetary value criterion is to be employed ?

Solution.

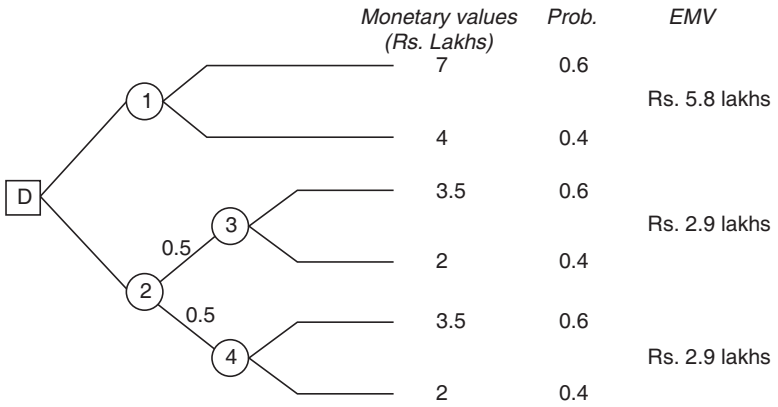


Thus, EMV for C is maximum, so alternative C is the best with maximum return of ₹ 4,900.

Example 3. *A manager has a choice between (i) a risky contract promising ₹ 7 lakhs with probability 0.6 and ₹ 4 lakhs with probability 0.4 (ii) a diversified portfolio consisting of*

two contracts with independent outcomes each promising ₹ 3.5 lakhs with probability of 0.6 and ₹ 2 lakhs with probability of 0.4. Construct a decision tree and suggest which choice the manager should opt using EMV criterion.

Solution.



EMV at node 2 = (2.9 × 0.5) + (2.9 × 0.5) = ₹ 2.9 lakhs.

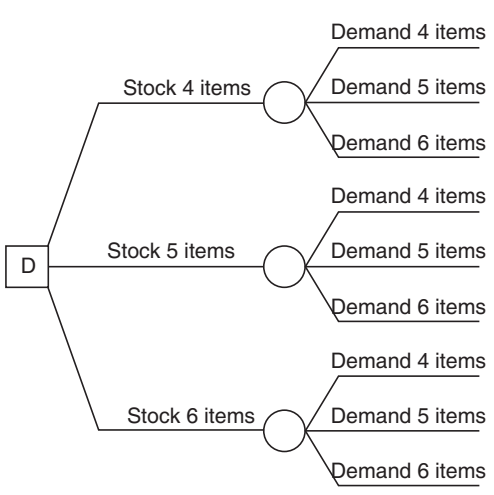
Thus, EMV for strategy (*i*) risky contract is maximum. So manager should opt choice (*i*).

Example 4. *A shopkeeper has the facility to store a large number of perishable items. He buys them at a rate of ₹ 3 per item and sells at the rate of ₹ 5 per item. If an item is not sold at the end of the day, then there is a loss of ₹ 3 per item. The daily demand of the item has the following probability distribution :*

Number of items sold	*4*	*5*	*6*
Probability	*0.2*	*0.5*	*0.3*

How many items should he store so that his daily expected profit is maximum ? Use decision tree approach.

Solution. Profit per item = ₹ (5 – 3) = ₹ 2



	Profit (Rs.)	*Prob.*	*EMV*
Stock 4 items, Demand 4 items	8	0.2	
Stock 4 items, Demand 5 items	8	0.5	Rs. 8
Stock 4 items, Demand 6 items	8	0.3	
Stock 5 items, Demand 4 items	5	0.2	
Stock 5 items, Demand 5 items	10	0.5	Rs. 9
Stock 5 items, Demand 6 items	10	0.3	
Stock 6 items, Demand 4 items	2	0.2	
Stock 6 items, Demand 5 items	7	0.5	Rs. 7.5
Stock 6 items, Demand 6 items	12	0.3	

Thus, EMV for strategy second is maximum. So shopkeeper should stock 5 items.

Example 5. *Matrix company is planning to launch a new product, which can be introduced initially in Western India or in the entire country. If the product is introduced only in Western India, the investment outlay will be ₹ 12 million. After two years, Matrix can evaluate the project to determine whether it should cover the entire country. For such expansion it will have to incur an additional investment of ₹ 10 million. To introduce the product in the entire country right in the begining would involve an outlay of ₹ 20 million. The product, in any case, will have a life of 5 years after which the plant will have zero net value.*

If the product is introduced only in Western India, demand would be high or low with the probabilities of 0.8 and 0.2 respectively and annual cashflow of ₹ 4 million and ₹ 2.5 million respectively.

If the product is introduced in the entire country right in the begining the demand would be high or low with probabilities of 0.6 and 0.4 respectively and annual cash inflows of ₹ 8 million and ₹ 5 million respectively.

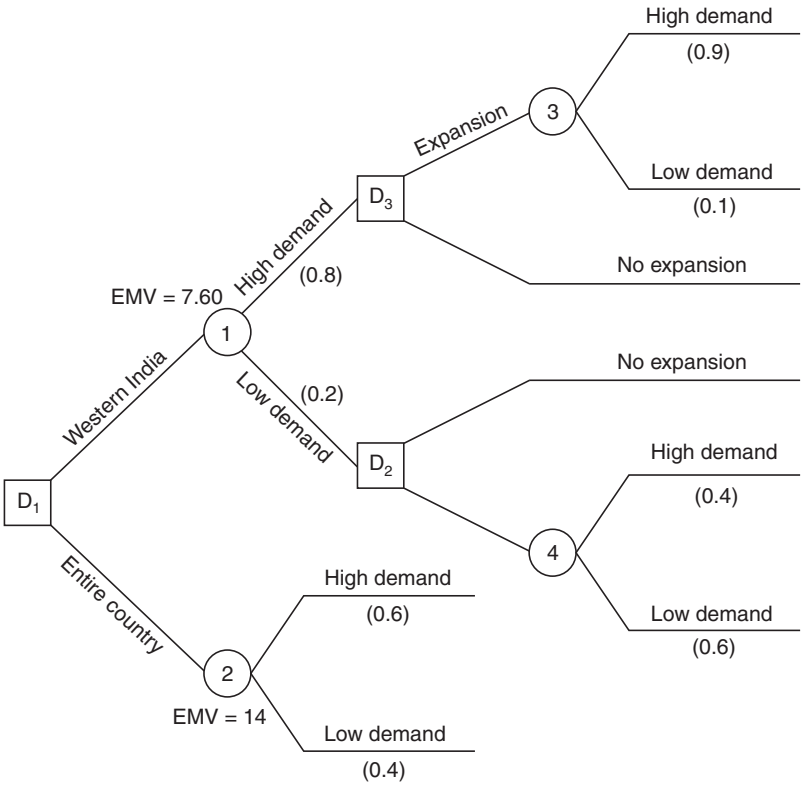
Based on the observed demand in Western India, if the product is introduced in the entire country the following probabilities would exist for high and low demand on an all India basis :

Western India	*Entire Country*	
	High demand	*Low demand*
High demand	*0.9*	*0.1*
Low demand	*0.4*	*0.6*

NOTES

The hurdle rate applicable to this project is 12 percent.
(i) Set up a decision tree for the investment situation.
(ii) Advice Matrix company on the investment policy it should follow. Support your advice with appropriate reasoning.

Solution.



	Decision Point	*Outcome*	*Probability*	*Conditional Value (₹)*	*Expected Value*
	(Demand high in Western India)				
	(*i*) Expansion	High demand	0.9	8	7.2
		Low demand	0.1	5	0.5
D3					7.7
					7.7 × 3 years = 23.1
					Less cost = 10.0
	(*ii*) No expansion				Total 13.1
					0
					Total expected profit = 13.1

	(Demand low in Western India)				
	(*i*) Expansion	High demand	0.4	8	3.2
		Low demand	0.6	5	3.0
					6.2
D2					6.2 × 3 years = 18.6
					Less cost = 10.0
	(*ii*) No expansion				Total 8.6
					0
					Total expected profit = 8.6
	(*i*) Introduction	High demand	0.6	8	4.8
	Entire country	Low demand	0.4	5	2.0
					6.8
					6.8 × 5 years = 34.0
					Less cost = 20.0
D1					Total expected profit 14
	(*ii*) Introduction	High demand	0.8	4	0.8 (4 × 2 + 13.1) = 16.88
	Western India	Low demand	0.2	2.5	0.2 (2.5 × 2 + 8.6) = 2.72
					= 19.60
					Less cost = 12.00
					Total expected profit = 7.6

Thus, the EMV at node 2 is maximum, make a decision to launch the product in entire country.

EXERCISE 3.1

1. The ABC company is faced with four decision alternatives relating to investments in a capital expansion programme. Since these investments are made in future, the company foresees different market conditions as expressed in the form of states of nature. The following table summarizes the decision alternatives, the various states of nature and the rate of return associated with each state of nature :

Decision	*States of nature*		
	Q1	*Q2*	*Q3*
D1	17%	15%	8%
D2	18%	16%	9%
D3	21%	14%	9%
D4	19%	12%	10%

If the company has no information regarding the probability of occurrence of the three states of nature, give the recommended decision for the decision criterion as follows :

(*i*) Maximax criterion (*ii*) Maximin criterion

(*iii*) Minimax regret criterion (*iv*) Laplace criterion

NOTES

2. A food products company is contemplating the introduction of a revolutionary new product with new packaging to replace the existing product at much higher price (A_1) or a moderate change in the composition of the existing product with a new packaging at a small increase in price (A_2) or a small change in the composition of the existing except the word 'New' with a negligible increase in price (A_3). The three possible states of nature or events are (*i*) high increase in sales (S_1), (*ii*) no change in sales (S_2) and (*iii*) decrease in sales (S_3). The marketing department of the company worked out the pay-offs in terms of yearly net profits for each of the strategies of three events (expected sales). This is represented in the following table :

Strategy	*States of nature Pay-offs (in ₹)*		
	S1	*S2*	*S3*
A1	7,00,000	3,00,000	1,50,000
A2	5,00,000	4,50,000	0
A3	3,00,000	3,00,000	3,00,000

Which strategy should the concerned executive choose on the basis of

(*i*) Maximin criterion (*ii*) Maximax criterion

(*iii*) Minimax regret criterion (*iv*) Laplace criterion ?

3. A person wants to invest in one of three alternative investment plans : stock, bonds or a savings account. It is assumed that the person wishes to invest all of the funds in one plan. The conditional pay-offs of the investments are based on three potential economic conditions : accelerated, normal or slow growth. The pay-off matrix is given as :

Alternative Investment	*Economic conditions Pay-off (₹)*		
	Accelerated growth	*Normal growth*	*Slow growth*
Stocks	10,000	6,500	– 4,000
Bonds	8,000	6,000	1,000
Savings	5,000	5,000	5,000

Determine the best investment plan using each of the following :

(*i*) Maximin criterion (*ii*) Maximax criterion

(*iii*) Laplace criterion (*iv*) Hurwicz criterion with $\alpha = 0.6$

4. Consider the following cost matrix :

Alternatives	*States of nature*			
	S1	*S2*	*S3*	*S4*
A1	1	3	8	5
A2	2	5	4	7
A3	4	6	6	3
A4	6	8	3	5

Determine the best alternative using :

(*i*) Minimax criterion

(*ii*) Minimin criterion

(*iii*) Minimax regret criterion.

5. Pay-offs (in ₹) of three acts A1, A2 and A3 and the possible states of nature S1, S2 and S3 are as follows :

State of nature	*Pay-offs (₹)*		
	Act		
	A1	*A2*	*A3*
S1	– 20	– 50	200
S2	200	– 100	– 50
S3	400	600	300

The probabilities of the states of nature are 0.3, 0.4 and 0.3 respectively. Tabulate the expected monetary values (EMVs) for the above data and state which can be selected as the best act.

6. An investor is given the following investment alternatives and percentage rates of return:

Strategy	*States of nature (Market conditions)*		
	Low	*Medium*	*High*
Regular shares	7%	10%	15%
Risky shares	– 10%	12%	25%
Property	– 12%	18%	30%

Over the past 300 days, 150 days have been medium market conditions and 60 days have been high market conditions.

On the basis of these data, state the optimum investment strategy for the investment.

7. Pay-offs (₹) of three acts A1, A2, A3 and the states of nature S1, S2 and S3 are as follows :

States of nature	*Pay-offs (₹)*		
	Act		
	A1	*A2*	*A3*
S1	25	– 10	– 125
S2	400	440	400
S3	650	740	750

The probabilities of the states of nature are 0.1, 0.7 and 0.2 respectively. Tabulate the expected monetary values (EMVs) and state which can be selected as the best act.

8. XYZ flower shop promises its customers delivery within four hours on all flower orders. All flowers are purchased on the prior day and delivered to XYZ by 8:00 the next morning. XYZ's daily demand for roses is as follows :

Dozens roses	7	8	9	10
Probability	0.1	0.2	0.4	0.3

XYZ purchases roses for ₹ 10.00 per dozen and sells them for ₹ 30.00. All unsold roses are donated to a local hospital. How many dozens of roses should XYZ order each evening to maximize its profit ? What is the optimum expected profit ?

9. A producer of boats has estimated the following distribution of demand for a particular kind of boat :

No. demanded	0	1	2	3	4	5	6
Probability	0.14	0.27	0.27	0.18	0.09	0.04	0.01

NOTES

Each boat cost him ₹ 7,000 and he sells them for ₹ 10,000 each. Any boat that are left unsold at the end of the season must be disposed off for ₹ 6,000 each. How many boats should be in stock so as to maximize his expected profit ?

10. Consider the following pay-off table.

Acts	*Events*			
	E1	*E2*	*E3*	*E4*
A1	18	10	12	8
A2	16	12	10	10
A3	12	13	11	12

The probabilities of events E1, E2, E3 and E4 are 0.25, 0.40, 0.15 and 0.20 respectively. Find the optimum act using expected opportunity loss (EOL) criterion.

11. A man has the choice of running either a hot-snack stall or an ice-cream stall at a seaside resort during the summer season. If it is a fairly cool summer, he should make ₹ 5,000 by running the hot-snack stall, but if the summer is quite hot he can only expect to make ₹ 1000. On the other hand, if he operates the ice-cream stall, his profit is estimated at ₹ 6500 if the summer is hot, but only ₹ 1000 if it is cool. There is a 40% chance of the summer being hot. Should he opt for running the hot-snack stall or the ice-cream stall ? Give mathematical argument.

12. The cost of making an item is ₹ 25, the selling price of the item is ₹ 30, if it is sold within a week, and it could be disposed off at ₹ 20 per piece at the end of week if unsold. Frequency of weekly sales is given as :

Weekly sales	(≤ 3)	4	5	6	7	(≥ 8)
No. of weeks	0	10	20	40	30	0

Find the optimum number of items per week the industry should make using EMV and EOL criterion. Also find, the EVPI.

13. A company wants to know whether or not a new shaving cream should be marketed. The present value of all future profits for the success of the cream is ₹ 10,00,000 and its failure would results in a net loss of ₹ 5,00,000.

Not marketing it would not change the profits. The chances of the success of the new cream are 50%. Determine the optimum act and find the EVPI.

14. A modern home appliances dealer finds that the cost of holding a mini-cooking range in stock for a month is ₹ 200 (insurance, minor deterioration, interest on borrowed capital, etc.). Customer who cannot obtain a cooking range immediately tends to go to other dealers and he estimates that for every customer who cannot get immediate delivery, he loses an average of ₹ 500. The probabilities of a demand of 0, 1, 2, 3, 4 and 5 cooking ranges in a month are 0.05, 0.10, 0.20, 0.30, 0.20 and 0.15 respectively. Determine the optimum stock level of cooking range. Also find the EVPI.

15. A manufacturer of leather goods must decide whether to expand his plant capacity now or wait at least another year. His advisors tell him that if he expands now and economic conditions remained good, there will be a profit of ₹ 1,64,000 during the next year. If he expands now and there is recession, there will be a loss of ₹ 40,000. If he waits at least another year and economic conditions remain good, there will be a profit of ₹ 80,000 and if he waits at least another year and there is a recession, there will be a small profit of ₹ 8,000. What should the manufacturer decide to do if he wants to minimize the expected loss during next year and he feels that the odds are 2 : 1 that there will be recession. Use decision tree approach.

16. XYZ Ltd. wants to update/change its existing manufacturing prices for product A. It wants to strengthen its research and development cell and conduct research for finding a better product of manufacturing, which can get them higher profits. At present the

company is earning a profit of ₹ 20,000 after paying for material, labour and overheads. XYZ Ltd. has the following four alternatives :

(*i*) The company continues with the existing process.

(*ii*) The company conducts research P, which costs ₹ 20,000, has 75% probability of success and can get the profit of ₹ 5,000.

(*iii*) The company conducts research Q, which costs ₹ 10,000, has 50% probability of success and can get the profit of ₹ 25,000.

(*iv*) The company pays ₹ 10,000 as royalty for a new product and can get profit of ₹ 20,000. The company can carry out only one out of the two types of research P and Q because of certain limitations. Draw a decision tree diagram and find the best strategy for XYZ Ltd.

17. The investment staff of a bank is considering four investment proposals for clients, shares, bonds, real estate and saving certificates, these investments will be held for one year. The past data regarding the four proposals is given as follows :

Shares. There is 25% chance that shares will decline by 10%, 30% chance that they will remain stable and 45% chance that they will increase in value by 15%. Also the shares under consideration do not pay any dividends.

Bonds. These bonds stand a 40% chance of increase in value by 5% and 60% chance of remaining stable and they yield 12%.

Real Estate. This proposal has a 20% chance of increasing 30% in value, a 25% chance of increasing 20% in value, a 40% chance of increasing 10% in value, 10% chance of remaining stable and a 5% chance of loosing 5% of its value.

Saving Certificates. These certificates will yields 8.5% with certainty.

Use a decision tree to structure the alternatives available to the investment staff, and using the expected monetary value criteria, choose the alternative with the highest expected value.

18. A manufacturing company has to select one of the two products A or B for manufacturing product A requires investment of ₹ 20,000 and product B ₹ 40,000. Market research survey shows high, medium and low demands with corresponding probabilities and return from sales, in ₹ thousand, for the two products, in the following table :

Market	*Probability*		*Return for sales*	
	A	*B*	*A*	*B*
High	0.4	0.3	50	80
Medium	0.3	0.5	30	60
Low	0.3	0.2	10	50

Construct an appropriate decision tree. What decision the company should take ?

Answers

1. (*i*) D3 (*ii*) D4 (*iii*) D3 (*iv*) D3 **2.** (*i*) A3 (*ii*) A1 (*iii*) A1 (*iv*) A1

3. (*i*) Savings (*ii*) Stock (*iii*) Bonds or Savings (*iv*) Bonds

4. (*i*) A3 (*ii*) A1 (*iii*) A3 **5.** A1 **6.** Property **7.** A2

8. 9 dozen, ₹ 168 **9.** 3 boats **10.** A2

11. Hot-snack stall **12.** 6 items, ₹ 3.50 **13.** Market cream, ₹ 2,50,000

14. 4 cooking ranges, ₹ 315. **15.** Wait for one year

16. Conduct research P to find a new process **17.** Invest in real estate **18.** Product B

NOTES

4. SAMPLING AND SAMPLING DISTRIBUTIONS

STRUCTURE

Sampling
Types of Sampling
Use of Random Numbers
Parameter and Statistic
Sampling Distribution of Mean
Sampling Distribution of Sample Variance
Sampling Distribution of Sample Proportion
Estimation
Point Estimation
Interval Estimation
Bayesian Estimation

4.1. SAMPLING

Sampling means the selection of a part of the aggregate with a view to draw some statistical informations about the whole. This aggregate of the investigation is called population and the selected part is called sample. A population is finite or infinite according to its size *i.e.,* number of members.

The main objective of the sampling is to obtain the maximum information of the population. The analysis of the sample is done to obtain an idea of the probability distribution of the variable in the population.

Though by applying proper process of sampling we may not be able to represent the characteristics of the population correctly. This discrepancy is called sampling error.

4.2. TYPES OF SAMPLING

There are different sampling methods. We describe below some important types of sampling.

(*a*) Simple random sampling. In this type of sampling every unit of the population has an equal chance of being selected in a sample. There are two ways of

drawing a simple random sample–With Replacement (WR) and Without Replacement (WOR).

In WR type, the drawn unit of the population is again returned to the population so that the size of the population remains same before each drawing. In WOR type, the drawn unit of the population is not returned to the population. For finite population the size diminishes as the sampling process continues.

(*b*) Systematic sampling. In systematic sampling one unit is chosen at random from the population and the items are selected regularly at predetermined intervals. This method is quite good over the simple random sampling provided there is no deliberate attempt to change the sequence of the units in the population.

(*c*) Cluster sampling. When the population consists of certain group of clusters of units, it may be advantageous and economical to select a few clusters of units and then examine all the units in the selected clusters. For example of certain goods which are packed in cartons and repacking is costly it is advisable to select only few cartons and inspect all the inside goods.

(*d*) Two-stage sampling. When the population consists of larger number of groups each consisting of a number of items, it may not be economical to select few groups and inspect all the items in the groups. In this case, the sample is selected in two stages. In the first stage, a desired number of groups (primary units) are selected at random and in the second stage, the required number of items are chosen at random from the selected primary units.

(*e*) Stratified sampling. Here the population is subdivided into several parts, called strata showing the heterogenity of the items is not so prominent and then a sub sample is selected from each of the strata. All the sub-samples combined together give the stratified sample. This sampling is useful when the population is heterogeneous.

4.3. USE OF RANDOM NUMBERS

The random numbers represent a sequence of digits where they appear in a perfectly random order. Selection of a random number from a table of random numbers has the same probability of selection. There are various methods to generate random numbers. Also there are tables of random numbers. Briefly we illustrate the use of random numbers. Let us consider the following two digits random numbers:

23, 04, 82, 07, 14, 66, 54, 10, 72 and 32.

Suppose we have marks of a subject of 100 students and we want to draw a sample of marks of size 10. To draw this, number the students from 00 to 99 and using the above random numbers select the marks of a student whose number is 23 since the first random is 23. Next select a student whose number is 04 since the next random number is 04. Repeating this process we obtain a sample of marks of size 10.

By considering another set of 10 random numbers, we can construct another sample of marks of size 10 and so on.

4.4. PARAMETER AND STATISTIC

Any statistical measure relating to the population which is based on all units of the population is called **parameter**, *e.g.,* population mean (μ), population S.D. (σ), moments μ_r, μ'_r etc.

NOTES

Any statistical measure relating to the sample which is based on all units of the sample is called **statistic,** *e.g.,* sample mean ($\bar{x}$), sample variance, moments m_r, m'_r etc. Hence the value of a statistic varies from sample to sample. This variation is called **'sampling fluctuation'.** The parameter has no fluctuation and it is constant. The probability distribution of a statistic is called 'sampling distribution'. The standard deviation (S.D.) in the sampling distribution is called **'standard error'** of the statistic.

Example 1. *For a population of five units, the values of a characteristic x are given below:*

8, 2, 6, 4 and 10.

Consider all possible samples of size 2 from the above population and show that the mean of the sample means is exactly equal to the population mean.

Solution. The population mean, $\mu = \frac{30}{5} = 6$

Random samples of size two (Without Replacement)

Serial no.	*Sample values*	*Sample mean*	*Serial no.*	*Sample values*	*Sample mean*
1	8, 2	5	6	2, 4	3
2	8, 6	7	7	2, 10	6
3	8, 4	6	8	6, 4	5
4	8, 10	9	9	6, 10	8
5	2, 6	4	10	4, 10	7
	Total	31		Total	29

$\therefore$ Mean of sample means $= \frac{31 + 29}{10} = \frac{60}{10} = 6$ which is equal to the population mean.

4.5. SAMPLING DISTRIBUTION OF MEAN

Case I : σ Known

Consider a population having mean μ and variance σ^2. If a random sample of size n is taken from this population then the sample mean $\bar{X}$ is a random variable whose distribution has the mean μ.

If the population is infinite, then the variance of this distribution is $\frac{\sigma^2}{n}$ and the standard error is defined as S.E. $= \frac{\sigma}{\sqrt{n}}$.

If the population is finite of size N then the variance of this distribution is $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$ and the standard error is defined as

$$\text{S.E.} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

provided the sample is drawn without replacement.

The factor $\frac{N-n}{N-1}$ is called finite population correction factor.

Let us consider the standardized sample mean

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

Then we have the central limit theorem as follows:

If $\bar{X}$ is the mean of a sample of size n taken from a population whose mean is μ and variance is σ^2, then

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0, 1) \text{ as } n \to \infty.$$

If the samples come from a normal population then the sampling distribution of the mean is normal regardless of the size of the sample.

If the population is not normal then the sampling distribution of the mean is approximately normal for small size ($n = 25$) of the sample.

Example 2. *A random sample of size 100 is taken from an infinite population having the mean $\mu = 66$ and the variance $\sigma^2 = 225$. What is the probability of getting an $\bar{x}$ between 64 and 68?*

Solution. Let $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, $n = 100$, $\mu = 66$, $\sigma = 15$

Required probability $= P[64 < \bar{x} < 68]$

$= P[-1.33 < z < 1.33]$

$= 2\Phi(1.33) = 2(0.4082)$

$= 0.8164.$

Example 3. *A random sample is of size 5 is drawn without replacement from a finite population consisting of 35 units. If the population standard deviation is 2.25. What is the standard error of sample mean?*

Solution. Here, $n = 5$, $N = 35$, $\sigma = 2.25$

S.E. of sample mean $= \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

$= \frac{2.25}{\sqrt{5}} \cdot \sqrt{\frac{30}{34}} = 0.95.$

Case II : σ Unknown

For small sample, the assumption of normal population gives fairly the sampling distribution of $\bar{X}$. However the σ is replaced by sample standard deviation S. Then we have

$$t = \frac{\bar{X}-\mu}{S/\sqrt{n}} \quad \text{where, } S^2 = \frac{1}{n-1} \cdot \Sigma(x_i - \bar{x})^2$$

is a random variable having the t distribution with the degrees of freedom $v = n - 1$.

NOTES

4.6. SAMPLING DISTRIBUTION OF SAMPLE VARIANCE

Like sample mean, if we calculate the sample variance for each samples drawn from a population then it shows also a random variable. We have the following result: If a random sample of size n with sample variance S^2 is taken from a normal population having the variance σ^2, then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad \text{where, } S^2 = \frac{1}{n-1} \cdot \Sigma(x_i - \bar{x})^2$$

is a random variable having the chi-square distribution with the degrees of freedom $v = n - 1$.

(In chi-square distribution table χ^2_α represents the area under the chi-square distribution to its right is equal to α).

If S_1^2 and S_2^2 are the variances of independent random sample of size n_1 and n_2 respectively, taken from two normal populations having the same variance, then

$$F = \frac{S_1^2}{S_2^2}$$

is a random variable having the F distribution with the degrees of freedoms $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$.

Example 4. *If two independent random samples of size $n_1 = 9$ and $n_2 = 16$ are taken from the normal population, what is the probability that the variance of the first sample will be at least four times as large as that of the second sample?*

Solution. Here $v_1 = 9 - 1 = 8$, $v_2 = 16 - 1 = 15$, $S_1^2 = 4S_2^2$

From F distribution table we find that

$$F_{0.01} = 4.00 \text{ for } v_1 = 8 \text{ and } v_2 = 15.$$

Thus, the desired probability is 0.01.

4.7. SAMPLING DISTRIBUTION OF SAMPLE PROPORTION

Consider a lot with proportion of defectives P. If a random sample of size n with proportion of defectives p is drawn from this population then the sampling distribution of p is approximately normal distribution with mean = P and S.D. = S.E. of sample proportion $= \sqrt{\frac{PQ}{n}}$ where, Q = 1 – P and the sample size n is sufficiently large. If the random sample is drawn from a finite population without replacement then we have to multiply a correction factor $\sqrt{\frac{N-n}{N-1}}$ to the S.D. formula.

If p_1 and p_2 denote the proportions from independent samples of sizes n_1 and n_2 drawn from two populations with proportions P_1 and P_2 respectively then

$$\text{S.E. of } (p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

where, $P_1 + Q_1 = 1$ and $P_2 + Q_2 = 1$.

NOTES

Example 5. *It has been found that 3% of the tools produced by a certain machine are defective. What is the probability that in a shipment of 450 such tools, 2% or more will be defective?*

Solution. Since the sample size $n = 450$ is large, the sample proportion (p) is approximately normally distributed with mean = P = 3% = 0.03.

$$\text{S.D.} = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{(0.03)\,(0.97)}{450}} = 0.008$$

$\therefore$ Required probability = $P[p > 0.02]$

$= P[z > -1.25] = 0.5 + \Phi\,(1.25)$

$= 0.5 + 0.3944 = 0.8944.$

EXERCISE 4.1

1. A population consists of 5 numbers (2, 3, 6, 8, 11). Consider all possible samples of size two which can be drawn with replacement from this population. Calculate the S.E. of sample means.
2. When we sample from an infinite population, what happens to the standard error of the mean if the sample size is (*a*) increased from 30 to 270, (*b*) decreased from 256 to 16?
3. A random sample of size 400 is taken from an infinite population having the mean μ = 86 and the variance of $\sigma^2 = 625$. What is the probability that $\overline{X}$ will be greater than 90?
4. The number of letters that a department receives each day can be modeled by a distribution having mean 25 and standard deviation 4. For a random sample of 30 days, what will be the probability that the sample mean will be less than 26?
5. A random sample of 400 mangoes was taken from a large consignment and 30 were found to be bad. Find the S.E. of the population of bad ones in a sample of this size.
6. From a population of large number of men with a S.D. 5, a sample is drawn and the standard error is found to be 0.5, what is the sample size?
7. A population consists of 20 elements, has mean 9 and S.D. 3 and a sample of 5 elements is taken without replacement. Find the mean and S.D. of the sampling distribution of the mean. What will be the S.D. for samples of size 10?
8. A machine produces a component for a transistor set of the total produce, 6 percent are defective. A random sample of 5 components is taken for examination from (*i*) a very large lot of produce, (*ii*) a box of 10 components. Find the mean and S.D. of the average number of defectives found among the 5 components taken for examination.
9. A population consists of five numbers 2, 3, 6, 8, 11. Consider all possible samples of size two which can be drawn without replacement from the population. Find
   (*a*) The mean of the population
   (*b*) Standard deviation of the population
   (*c*) The mean of the sampling distribution of means
   (*d*) The standard deviation of the sampling distribution of means.

Answers

1. 2.32 **2.** (*a*) It is divided by 3 (*b*) It is multiplied by 4

3. 0.0007 **4.** 0.9147

5. 0.013 **6.** 100

NOTES

7. For sample of 5 elements, sampling mean = 8, S.D. = $\sqrt{\frac{27}{19}}$

For sample of 10 elements, sampling mean = 8, S.D. = $\frac{3}{\sqrt{19}}$

8. Mean = 0.06, S.D. = 0.106

9. (*a*) 6, (*b*) 3.29, (*c*) 6, (*d*) 2.12.

4.8. ESTIMATION

When we deal with a population, most of the time the parameters are unknown. So we cannot draw any conclusion about the population. To know the unknown parameters the technique is to draw a sample from the population and try to gather information about the parameter through a function which is reasonably close. Thus the obtained value is called an estimated value of the parameter, the process is called estimation and the estimating function is called estimator.

A good estimator should satisfy the four properties which we briefly explain below:

(*a*) **Unbiasedness.** A statistic t is said to be an unbiased estimator of a parameter θ if, $E[t] = \theta$.

Otherwise it is said to be 'biased'.

Theorem 1. Prove that the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ.

Proof. Let $x_1, x_2, ..., x_n$ be a simple random sample with replacement from a finite population of size N, say, $X_1, X_2, ..., X_N$

Here, $\bar{x} = (x_1 + x_2 + ... + x_n)/n$

$\mu = (X_1 + X_2 + ... + X_N)/N$

To prove that $E(\bar{x}) = \mu$

While drawing x_i, it can be one of the population members *i.e.*, the probability distribution of x_i can be taken as follows:

x_i	X_1	X_2	$...X_N$
Probability	1/N	1/N	1/N

for $i = 1, 2, ..., n$

Therefore,

$$E(x_i) = X_1 \cdot \frac{1}{N} + X_2 \cdot \frac{1}{N} + ... + X_N \cdot \frac{1}{N}$$

$$= (X_1 + X_2 + ... + X_N)/N$$

$$= \mu, \ i = 1, 2, ..., n.$$

and

$$E(\bar{x}) = E[(x_1 + x_2 + ... + x_n)/n]$$

$$= [E(x_1) + E(x_2) + ... + E(x_n)]/n$$

$$= [\mu + \mu + ... + \mu]/n = n\mu/n = \mu.$$

The same result is also true for infinite population and the sampling without replacement.

Theorem 2. The sample variance

$$S^2 = \frac{1}{n} \cdot \Sigma (x_i - \bar{x})^2$$

is a biased estimator of the population variance σ^2.

Proof. Let $x_1, x_2, ..., x_n$ be a random sample from an infinite population with mean μ and variance σ^2.

Then $E(x_i) = \mu$, $Var(x_i) = E(x_i - \mu)^2 = \sigma^2$, for $i = 1, 2, ..., n$.

$$s^2 = \frac{1}{n} \cdot \Sigma (x_i - \bar{x})^2$$

$$= \frac{1}{n} \cdot \Sigma x_i^2 - (\bar{x})^2$$

$$= \frac{1}{n} \cdot \Sigma y_i^2 - (\bar{y})^2, \quad \text{where, } y_i = x_i - \mu \text{ and S.D is unaffected by change of origin.}$$

$$= \frac{1}{n} \cdot \Sigma (x_i - \mu)^2 - (\bar{x} - \mu)^2$$

$\therefore$ $$E(s^2) = \frac{1}{n} \cdot \Sigma E(x_i - \mu)^2 - E(\bar{x} - \mu)^2$$

$$= \frac{1}{n} .\Sigma \sigma^2 - Var(\bar{x}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \cdot \sigma^2 \neq \sigma^2.$$

$\Rightarrow$ s^2 is a biased estimator of σ^2

Note. Let $S^2 = \frac{1}{(n-1)} \cdot \Sigma (x_i - \bar{x})^2$, then

$$E(S^2) = \frac{n}{n-1} .E(s^2)$$

$$= \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 = \sigma^2$$

Thus, S^2 is an unbiased estimator of σ^2.

Example 1. *A population consists of 4 values 3, 7, 11, 15. Draw all possible sample of size two with replacement. Verify that the sample mean is an unbiased estimator of the population mean.*

Solution. No. of samples = $4^2 = 16$, which are listed below:

(3, 3),	(7, 3),	(11, 3),	(15, 3)
(3, 7),	(7, 7),	(11, 7),	(15, 7)
(3, 11),	(7, 11),	(11, 11),	(15, 11)
(3, 15),	(7, 15),	(11, 15),	(15, 15)

NOTES

$$\text{Population mean, } \quad \mu = \frac{3+7+11+15}{4} = \frac{36}{4} = 9$$

Sampling distribution of sample mean

Sample mean $(\bar{x})$	*Frequency* $f(\bar{x})$	$\bar{x} \cdot f(\bar{x})$
3	1	3
5	2	10
7	3	21
9	4	36
11	3	33
13	2	26
15	1	15
Total	16	144

$$\text{Mean of sample mean} = \frac{144}{16} = 9$$

Since, $\quad E(\bar{x}) = \mu,$

$\Rightarrow$ Sample mean is an unbiased estimator of the population mean.

(*b*) **Consistency.** A statistic t_n obtained from a random sample of size n is said to be a consistent estimator of a parameter if it converges in probability to θ as n tends to infinity.

Alt, If $E[T_n] \to \theta$ and $\text{Var}[T_n] \to 0$ as $n \to \infty$, then the statistic t_n is said to be consistent estimator of θ.

For example, in sampling from a Normal Population $N(\mu, \sigma^2)$,

$$E[\bar{x}] = \mu \text{ and } \text{Var}[\bar{x}] = \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty.$$

Hence, the sample mean is a consistent estimator of population mean.

(*c*) **Efficiency.** There may exist more than one consistent estimator of a parameter. Let T_1 and T_2 be two consistent estimators of a parameter θ. If Var (T_1) < Var (T_2) for all n then T_1 is said to be more efficient than T_2 for all sample size.

If a consistent estimator has least variance than any other consistent estimators of a parameter, then it is called the most efficient estimator.

Let T be the most efficient estimator and T_1 be any other consistent estimator of a parameter. Then, we define

$$\text{Efficiency} = \frac{\text{Var}(T)}{\text{Var}(T_1)}$$

which is less than equal to one.

NOTES

A statistic which is unbiased and also the most efficient, is said to be the Minimum Variance Unbiased Estimator (MVUE).

Note. If T_1 and T_2 are two MVU Estimators of a parameter then $T_1 = T_2$.

For example, the sample mean $\bar{x}$ obtained from a normal population is the MVUE for the parameter μ.

Let $x_1, x_2, ..., x_n$ be a random sample and

$$T = a_1x_1 + a_2x_2 + ... + a_nx_n$$

where $a_1, a_2, ..., a_n$ are constants. If T is an MVUE, then T is also called Best Linear Unbiased Estimator (BLUE).

Example 2. *A random sample $(X_1, X_2, X_3, X_4, X_5, X_6)$ of size 6 is drawn from a normal population with unknown mean μ. Consider the following estimators to estimate μ.*

(i) $$T_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}$$

(ii) $$T_2 = \frac{X_1 + X_2 + X_3}{2} + \frac{X_4 + X_5 + X_6}{3}$$

(iii) $$T_3 = \frac{1}{2}(X_1 + X_2) + X_3 + X_4 + \frac{1}{3}(X_5 + X_6)$$

Are these estimators unbiased? Find the estimator which is best among T_1, T_2 and T_3.

Solution. Here $E(X_i) = \mu$, $Var(X_i = \sigma^2$ (say), $Cov(X_i, X_j) = 0$, $i \neq j$

$$E(T_1) = \frac{1}{6}[E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) + E(X_6)]$$

$$= \frac{1}{6}[\mu + \mu + \mu + \mu + \mu + \mu) = \frac{1}{6}.6\,\mu = \mu.$$

$$E(T_2) = \frac{1}{2}[E(X_1) + E(X_2) + E(X_3)] + \frac{1}{3}[E(X_4) + E(X_5) + E(X_6)]$$

$$= \frac{1}{2}[\mu + \mu + \mu] + \frac{1}{3}[\mu + \mu + \mu] = \frac{3\mu}{2} + \mu = \frac{5\mu}{2}.$$

$$E(T_3) = \frac{1}{2}[E(X_1) + E(X_2)] + E(X_3) + E(X_4) + \frac{1}{3}[E(X_5) + E(X_6)]$$

$$= \frac{1}{2}[\mu + \mu] + \mu + \mu + \frac{1}{3}[\mu + \mu]$$

$$= \mu + 2\mu + \frac{2\mu}{3} = \frac{11\mu}{3}.$$

NOTES

Since $E(T_1) = \mu \Rightarrow T_1$ is unbiased. T_2 and T_3 are biased estimators.

$$\text{Var}(T_1) = \frac{1}{36}[\text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_6)]$$

$$= \frac{1}{36}[\sigma^2 + \sigma^2 + \ldots + \sigma^2] = \frac{1}{36}(6\sigma^2) = \frac{\sigma^2}{6}.$$

$$\text{Var}(T_2) = \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)] + \frac{1}{9}[\text{Var}(X_4) + \text{Var}(X_5) + \frac{1}{9}\text{Var}(X_6)]$$

$$= \frac{1}{4}[\sigma^2 + \sigma^2 + \sigma^2] + [\sigma^2 + \sigma^2 + \sigma^2]$$

$$= \frac{3}{4}\sigma^2 + \frac{3\sigma^2}{9} = \frac{13}{12}\sigma^2.$$

$$\text{Var}(T_3) = \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2)] + \text{Var}(X_3) + \text{Var}(X_4) + \frac{1}{9}[\text{Var}(X_5) + \text{Var}(X_6)]$$

$$= \frac{1}{4}[\sigma^2 + \sigma^2] + \sigma^2 + \sigma^2 + \frac{1}{9}[\sigma^2 + \sigma^2]$$

$$= \frac{\sigma^2}{2} + 2\sigma^2 + \frac{2\sigma^2}{9} = \frac{49}{18}\sigma^2.$$

Since $\text{Var}(T_1)$ is smallest $\Rightarrow T_1$ is best estimator.

$$\text{Efficiency of } T_1 \text{ over } T_2 = \frac{\sigma^2/6}{13\ \sigma^2/12} = \frac{2}{13} = 0.15$$

$$\text{Efficiency of } T_1 \text{ over } T_3 = \frac{\sigma^2/6}{49\ \sigma^2/18} = \frac{3}{49} = 0.06.$$

Example 3. *A random sample (X_1, X_2, X_3, X_4) of size 4 is drawn from a normal population with unknown mean. If*

$$T = 2X_1 + \frac{\lambda}{2}X_2 + 3X_3 - 4X_4$$

be an unbiased estimator of μ, find λ.

Solution. Let $E(X_i) = \mu$, $i = 1, 2, 3, 4$.

For unbiasedness, $E(T) = \mu$

$$\Rightarrow \quad 2E(X_1) + \frac{\lambda}{2}E(X_2) + 3E(X_3) - 4E(X_4) = \mu$$

$$\Rightarrow \quad 2\mu + \frac{\lambda}{2}\mu + 3\mu - 4\mu = \mu$$

$$\Rightarrow \quad \mu + \frac{\lambda}{2}\mu = \mu$$

$$\Rightarrow \quad \frac{\lambda}{2} = 0 \quad \Rightarrow \quad \lambda = 0.$$

(*d*) **Sufficiency.** Let $x_1, x_2, ..., x_n$ be a random sample from a population whose *p.m.f.* or *pdf* is $f(x, \theta)$. Then T is said to be a sufficient estimator of θ if we can express the following:

$$f(x_1, \theta) \cdot f(x_2, \theta) \ldots f(x_n, \theta) = g_1(T, \theta) \cdot g_2(x_1, x_2, ..., x_n)$$

where $g_1(T, \theta)$ is the sampling distribution of T and contains θ and $g_2(x_1, x_2, ..., x_n)$ is independent of θ.

Sufficient estimators exist only in few cases. However in random sampling from a normal population, the sampling mean $\bar{x}$ is a sufficient estimator of μ.

4.9. POINT ESTIMATION

Using sampling if a single value is estimated for the unknown parameter of the population, then this process of estimation is called point estimation. We shall discuss two methods of point estimation below:

I. Method of Maximum Likelihood

Let $x_1, x_2, ..., x_n$ be a random sample from a population whose *p.m.f.* (discrete case) or *p.d.f.* (continuous case) is $f(x, \theta)$ where θ is the parameter. Then construct the likelihood function as follows:

$$L = f(x_1, \theta). \quad f(x_2, \theta) \ldots f(x_n, \theta).$$

Since, log L is maximum when L is maximum. Therefore to obtain the estimate of θ, we maximize L as follows:

$$\frac{\partial}{\partial \theta} (\log L) = 0 \Rightarrow \theta = \hat{\theta}$$

and

$$\frac{\partial^2}{\partial \theta^2} (\log L) < 0 \text{ at } \theta = \hat{\theta}$$

Here $\hat{\theta}$ is called Maximum Likelihood Estimator (MLE).

Properties of MLE

(*i*) MLE is not necessarily unbiased.

(*ii*) MLE is consistent, most efficient and also sufficient, provided a sufficient estimator exists.

(*iii*) MLE tends to be distributed normally for large samples.

(*iv*) If $g(\theta)$ is a function of θ and $\hat{\theta}$ is an MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Example 4. *A discrete random variable X can take up all non-negative integers and*

$$P(X = r) = p(1 - p)^r \quad (r = 0, 1, 2, ...)$$

where, p (0 < p < 1) is the parameter of the distribution. Find the MLE of p for a sample of size n : $x_1, x_2, ..., x_n$ from the population of X.

Solution. Consider the following likelihood function:

$$L = P(X = x_1) \cdot P(X = x_2) \ldots P(X = x_n)$$

$$= p(1 - p)^{x_1} \cdot p(1 - p)^{x_2} \ldots p(1 - p)^{x_n}$$

$$= p^n (1 - p)^{x_1 + x_2 + \ldots + x_n} = p^n (1 - p)^{\Sigma x_i}$$

NOTES

Taking log on both sides we obtain

$$\ln L = n \ln p + (\Sigma x_i) \ln (1-p)$$

Now $$\frac{d \ln L}{dp} = 0$$

$$\Rightarrow \quad \frac{n}{p} - \frac{\Sigma x_i}{1-p} = 0$$

$$\Rightarrow \quad \frac{n}{p} = \frac{\Sigma x_i}{1-p}$$

$$\Rightarrow \quad \frac{1-p}{p} = \frac{\Sigma x_i}{n}$$

$$\Rightarrow \quad \frac{1}{p} - 1 = \bar{x}$$

$$\Rightarrow \quad \hat{p} = \frac{1}{1+\bar{x}}$$

Also,

$$\frac{d^2 \ln L}{dp^2} = -\frac{n}{p^2} - \frac{\Sigma x_i}{(1-p)^2} = -n\left(\frac{1}{p^2} + \frac{\bar{x}}{(1-p)^2}\right)$$

$$= -n\left((1+\bar{x})^2 + \frac{\bar{x}\,(1+\bar{x})^2}{(\bar{x})^2}\right)$$

at $$\hat{p} = \frac{1}{1+\bar{x}} = -n\,(1+\bar{x})^2\,(1+\frac{1}{x}) < 0$$

Hence the MLE of p is $\frac{1}{1+\bar{x}}$.

Example 5. *A random variable X has a distribution with density function:*

$$f(x) = \lambda\ x^{\lambda-1}\ (0 < x < 1)$$

where λ is the parameter. Find the MLE of λ for a sample of size n : $x_1, x_2, ..., x_n$ from the population of X.

Solution. Consider the following likelihood function:

$$L = f(x_1).\ f(x_2) \ldots f(x_n)$$

$$= \lambda\, x_1^{\lambda-1} .\ \lambda\, x_2^{\lambda-1} \ldots \lambda\, x_n^{\lambda-1}$$

$$= \lambda^n\, (x_1 .\ x_2 \ldots . x_n)^{\lambda-1}$$

Taking log on both sides we obtain

$$\ln L = n \ln \lambda + (\lambda - 1) \ln (x_1 .\ x_2 \ldots .\ x_n)$$

Now, $$\frac{d \ln L}{d \lambda} = 0 \quad \Rightarrow \quad \frac{n}{\lambda} + \ln (x_1\, x_2 \ldots x_n) = 0$$

$$\Rightarrow \quad \frac{n}{\lambda} = -\ln (x_1\, x_2 \ldots x_n)$$

$$\Rightarrow \quad \hat{\lambda} = \frac{-n}{\ln (x_1\, x_2 \ldots x_n)}$$

Also, $$\frac{d^2 \ln L}{d\lambda^2} = -\frac{n}{\lambda^2} < 0$$

Hence, the MLE of λ is $\frac{-n}{\ln(x_1 x_2 \ldots x_n)}$.

NOTES

Example 6. *X tossed a biased coin 40 times and got head 15 times, while Y tossed it 50 times and got head 30 times. Find the MLE of the probability of getting head when the coin is tossed.*

Solution. Let P be the unknown probability of getting a head.

Using binomial distribution,

Probability of getting 15 heads in 40 tosses = $\binom{40}{15} P^{15} (1 - P)^{25}$

Probability of getting 30 heads in 50 tosses = $\binom{50}{30} P^{30} (1 - P)^{20}$

The likelihood function is taken by multiplying these probabilities.

$$L = \binom{40}{15} \cdot \binom{50}{30} P^{45} (1 - P)^{45}$$

$$\therefore \quad \log L = \log \left[\binom{40}{15} \cdot \binom{50}{30} \right] + 45 \log P + 45 \log (1 - P)$$

Hence, $\frac{\partial \log L}{\partial P} = 0 \Rightarrow \frac{45}{P} - \frac{45}{1 - P} = 0 \Rightarrow P = 1/2$, which is the MLE.

II. Method of Moments

In this method, the first few moments of the population is equated with the corresponding moments of the sample.

Then $\mu'_r = m'_r$

where $\mu'_r = E(x^r)$ and $m'_r = \Sigma x_i^r / n$

The solution for the parameters gives the estimates. But this method is applicable only when the population moments exist.

Example 7. *Estimate the parameter p of the binomial distribution by the method of moments (when n is known).*

Solution. Here, $\mu'_1 = E(x) = np$ and $m'_1 = \bar{x}$

Taking $\mu'_1 = m'_1$, we have

$$np = \bar{x}$$

$$\Rightarrow \quad p = \frac{\bar{x}}{n}$$

which is the estimated value.

4.10. INTERVAL ESTIMATION

NOTES

In interval estimation we find an interval which is expected to include the unknown parameter with a specified probability, *i.e.,*

$$P(t_1 \le \theta \le t_2) = k$$

where, $[t_1, t_2]$ is called confidence interval (C.I.),

t_1, t_2 are called confidence limits,

k is called confidence co-efficient of the interval.

(*a*) C.I. for mean with known S.D. Let us consider a random sample of size n from a Normal Population N (μ, σ^2) **in which σ^2 is known.** To find C.I. for mean μ.

We know that $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ follows standard normal distribution and 95% of the area under the standard normal curve lies between $z = 1.96$ and $z = -1.96$, Then,

$$P[-1.96 \le z \le 1.96] = 0.95$$

$$\Rightarrow \quad P\left[-1.96 \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le 1.96\right] = 0.95$$

i.e., in 95% cases we have

$$-1.96 \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le 1.96$$

$$\Rightarrow \quad \bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}$$

The interval $\left[\bar{x} - 1.96\,\dfrac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\,\dfrac{\sigma}{\sqrt{n}}\right]$ is known as 95% confidence interval for μ.

Similarly, $\left[\bar{x} - 2.58\,\dfrac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\,\dfrac{\sigma}{\sqrt{n}}\right]$ is known as 99% C.I. for μ,

$\left[\bar{x} - 3.\,\dfrac{\sigma}{\sqrt{n}}, \bar{x} + \dfrac{3\sigma}{\sqrt{n}}\right]$ is known as 99.73% C.I. for μ.

(*b*) C.I. for mean with unknown S.D. σ.

In this case, the sampling from a normal population N (μ, σ^2), the statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}, \quad \text{where } s^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

follows t distribution with $(n - 1)$ degree of freedom.

Then for 95% confidence interval for mean μ we have

$$-t_{0.025} \le \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \le t_{0.025}$$

$$\Rightarrow \quad \bar{x} - t_{0.025}\cdot\frac{s}{\sqrt{n-1}} \le \mu \le \bar{x} + t_{0.025}\,\frac{s}{\sqrt{n-1}}$$

Thus, $\left[\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n-1}}, \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n-1}}\right]$ is called 95% C.I. for μ.

Similarly, $\left[\bar{x} - t_{0.005} \cdot \frac{s}{\sqrt{n-1}}, \bar{x} + t_{0.005} \cdot \frac{s}{\sqrt{n-1}}\right]$ is called 99% C.I. for μ.

NOTES

(*c*) **C.I. for variance σ^2 with known mean.** We know that $(x_i - \mu)^2/\sigma^2$ follows chi-square distribution with n degrees of freedom.

For probability 95% we have

$$\chi^2_{0.975} \leq \Sigma (x_i - \mu)^2 / \sigma^2 \leq \chi^2_{0.025}$$

$$\Rightarrow \quad \Sigma (x_i - \mu)^2 / \chi^2_{0.025} \leq \sigma^2 \leq \Sigma (x_i - \mu)^2 / \chi^2_{0.975}$$

which is 95% confidence interval for σ^2.

Similarly,

$$\Sigma (x_i - \mu)^2 / \chi^2_{0.005} \leq \sigma^2 \leq \Sigma (x_i - \mu)^2 / \chi^2_{0.995}$$

is the 99% confidence interval for σ.

(*d*) **C.I. for variance σ^2 with unknown mean.** In this case $ns^2 / \sigma^2 = \Sigma (x_i - \bar{x})^2 / \sigma^2$ follows chi-square distribution with $(n - 1)$ degrees of freedom.

For probability 95% we have

$$\chi^2_{0.975} \leq ns^2 / \sigma^2 \leq \chi^2_{0.025}$$

$$\Rightarrow \quad ns^2/\chi^2_{0.025} \leq \sigma^2 \leq ns^2/\chi^2_{0.975}$$

which is 95% C.I. for σ^2

Similarly, $ns^2 / \chi^2_{0.005} \leq \sigma^2 \leq ns^2 / \chi^2_{0.995}$ is the 99% C.I. for σ^2.

Some of the Confidence Limits are given below:

(with Normal Population N (μ, σ^2))

Difference of Means $(\mu_1 - \mu_2)$: (S.Ds known).

$$95\% \text{ Confidence limits} = (\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

$$99\% \text{ Confidence limits} = (\bar{x}_1 - \bar{x}_2) \pm 2.58 \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Difference of Means $(\mu_1 - \mu_2)$: (Common S.D. unknown)

$$95\% \text{ Confidence limits} = (\bar{x}_1 - \bar{x}_2) \pm t_{0.025} \cdot s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$99\% \text{ Confidence limits} = (\bar{x}_1 - \bar{x}_2) \pm t_{0.005} \cdot s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

For Proportion P:

95% Confidence limits = $p \pm 1.96$ (S.E. of p)

99% Confidence limits = $p \pm 2.58$ (S.E. of p)

where, S.E. of $p = \sqrt{\frac{PQ}{n}} \approx \sqrt{\frac{pq}{n}}$

NOTES

For Difference of Proportions $P_1 - P_2$:

95% Confidence limits = $(p_1 - p_2) \pm 1.96$ [S.E. of $(p_1 - p_2)$]

99% Confidence limits = $(p_1 - p_2) \pm 2.58$ [S.E. of $(p_1 - p_2)$]

where $$\text{S.E of } (p_1 - p_2) = \sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}} \approx \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}.$$

Example 8. *A random sample of size 10 was drawn from a normal population with an unknown mean and a variance of 35.4 (cm)². If the observations are (in cms): 55, 75, 71, 66, 73, 77, 63, 67, 60 and 76, obtain 99% confidence interval for the population mean.*

Solution. Given $n = 10$, $\Sigma x_i = 683$, Then $\bar{x} = \frac{\Sigma x}{n} = 68.3$

Since, the population S.D. σ is known, then 99% C.I. for μ is given by

$$\left[\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}\right]$$

i.e., $$\left[68.3 - \frac{2.58 \cdot \sqrt{35.4}}{\sqrt{10}}, 68.3 + \frac{2.58 \cdot \sqrt{35.4}}{\sqrt{10}}\right]$$

i.e., [63.45, 73.15].

Example 9. *A random sample of size 10 was drawn from a normal population which are given by 48, 56, 50, 55, 49, 45, 55, 54, 47, 43. Find 95% confidence interval for mean μ of the population.*

Solution. From the given data, $\Sigma x_i = 502$, so $\bar{x} = 50.2$, $n = 10$

Let $d = x - 50$, then the samples are changed to

$-2, 6, 0, 5, -1, -5, 5, 4, -3, -7$.

$\Sigma d = 2$, $\Sigma d^2 = 190$

$$\therefore \quad s^2 = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2 = \frac{190}{10} - \left(\frac{2}{10}\right)^2 = 18.96$$

$$s = 4.35$$

Since, the population S.D. σ is unknown, the 95% C.I. for mean μ is

$$\left[\bar{x} - 2.262 \cdot \frac{s}{\sqrt{n}}, \bar{x} + 2.262 \cdot \frac{s}{\sqrt{n}}\right]$$

i.e., $$\left[50.2 - (2.262) \frac{(4.35)}{\sqrt{10}}, 50.2 + (2.262) \frac{(4.35)}{\sqrt{10}}\right]$$

i.e., [47.09, 53.31].

Example 10. *The standard deviation of a random sample of size 15 drawn from a normal population is 3.2. Calculate the 95% confidence interval for the standard deviation (σ) in the population.*

Solution. Here $n = 15$, sample s.d. $(s) = 3.2$

95% Confidence interval for σ^2 is

$$\frac{n s^2}{\chi^2_{0.025}} \le \sigma^2 \le \frac{n s^2}{\chi^2_{0.975}}$$

NOTES

From chi-square table with 14 degrees of freedom,

$$\chi^2_{0.025} = 26.12, \quad \chi^2_{0.975} = 5.63$$

Therefore the C.I. is

$$\frac{15.(3.2)^2}{26.12} \leq \sigma^2 \leq \frac{15.(3.2)^2}{5.63}$$

i.e., $5.88 \leq \sigma^2 \leq 27.28$

i.e., $2.42 \leq \sigma \leq 5.22.$

Example 11. *A sample of 500 springs produced in a factory is taken from a large consignment and 65 are found to be defective. Estimate the assign limits in which the percentage of defectives lies.*

Solution. There are 65 defective springs in a sample of size $n = 500$.

$\therefore$ The sample proportion of defective is

$$p = \frac{65}{500} = 0.13$$

The limits to the percentage of defectives refer to the C.I., which can be taken as

$$[p - 3 \text{ (S.E. of } p), \; p + 3 \text{ (S.E. of } p)]$$

Here $$\text{S.E. of } p = \sqrt{\frac{PQ}{n}}$$

$$\approx \sqrt{\frac{pq}{n}} = \sqrt{\frac{65}{500}\left(1 - \frac{65}{500}\right) \cdot \frac{1}{500}} = 0.02$$

Thus, the limits are $[0.13 - 3\ (0.02),\ 0.13 + 3\ (0.02)]$

i.e., $[0.07, 0.19]$.

4.11. BAYESIAN ESTIMATION

Bayesian estimation uses subjective judgement in an engineering design.

For discrete case, let the parameter θ takes the values θ_i, $i = 1, 2, \ldots, n$ with the probabilities $p_i = P\ [\theta = \theta_i]$. Let θ_0 be the observed outcome of the experiment. Then by Bayes' theorem we obtain,

$$P[\theta = \theta_i \mid \theta_0] = \frac{P\ [\theta_0 \mid \theta = \theta_i] \,.\, P\ [\theta = \theta_i]}{\sum_{j=1}^{n} P\ [\theta_0 \mid \theta = \theta_j] \,.\, P\ [\theta = \theta_j]} \quad i = 1, 2, \ldots, n$$

Then the expected value of θ is called Bayesian estimator of the parameter, *i.e.*,

$$\hat{\theta} = E\left[\theta = \theta_i \mid \theta_0\right]$$

$$= \sum_{i=1}^{n} \theta_i \,.\, P\left[\theta = \theta_i \mid \theta_0\right]$$

Using this we can calculate

$$P\ [X \leq a] = \sum_{i=1}^{n} P\left[X \leq a \mid \theta = \theta_i\right] . P\left[\theta = \theta_i \mid \theta_0\right]$$

For continuous case, let θ be a random variable of the parameter of the distribution given by the density function $f'(\theta)$. Then

$$P[\theta_i < \theta < \theta_i + \Delta\theta] = f'(\theta_i) \,.\, \Delta\theta, \; i = 1, 2, \ldots, n$$

NOTES

If θ_0 is an observed experimental outcome, then

$$f''(\theta_i)\,\Delta\theta = \frac{P\left[\theta_0|\theta_i\right] . f'(\theta_i)\Delta\theta}{\sum_{j=1}^{n} P\left[\theta_0|\theta_j\right] f'\left(\theta_j\right)\Delta\theta}, \quad i = 1, 2, \ldots, n$$

In the limit we obtain, $f''(\theta) = \dfrac{P\left[\theta_0|\theta\right] f'(\theta)}{\int_{-\infty}^{\infty} P\left[\theta_0|\theta\right] f'(\theta)\, d\theta}$

Then the Bayesian estimator is

$$\hat{\theta} = E\left[\theta|\theta_0\right] = \int_{-\infty}^{\infty} \theta\, f''(\theta) d\theta$$

Using this we can calculate

$$P\left[X \le a\right] = \int_{-\infty}^{\infty} P\left[X \le a|\theta\right] f''(\theta) d\theta.$$

SUMMARY

- Sampling means the selection of a part of the aggregate with a view to draw some statistical informations about the whole. This aggregate of the investigation is called population and the selected part is called sample.
- Any statistical measure relating to the population which is based on all units of the population is called parameter.
- Any statistical measure relating to the sample which is based on all units of the sample is called statistic.
- When we deal with a population, most of the time the parameters are unknown. So we cannot draw any conclusion about the population. To know the unknown parameters the technique is to draw a sample from the population and try to gather information about the parameter through a function which is reasonably close. Thus, the obtained value is called an estimated value of the parameter, the process is called estimation and the estimating function is called estimator.
- If a consistent estimator has least variance than any other consistent estimators of a parameter, then it is called the most efficient estimator.
- Using sampling if a single value is estimated for the unknown parameter of the population, then this process of estimation is called point estimation.

EXERCISE 4.2

1. A random variable X has a distribution with density function:

$$f(x) = .(\alpha + 1)\, x^{\alpha}, \qquad 0 \le x \le 1, \alpha > -1$$
$$= 0, \qquad \text{otherwise}$$

and a random sample of size 8 produces the data: 0.2, 0.4, 0.8, 0.5, 0.7, 0.9, 0.8 and 0.9. Find the MLE of the unknown parameter α.

NOTES

2. A random variable X has a distribution with density function:

$$f(x) = \frac{(a+1)\,x^a}{2^{a+1}}, \qquad 0 \le x \le 2$$

$$= 0, \qquad \text{otherwise}$$

Find the MLE of the parameter a (> 0).

3. Consider a random sample of size n from a population following Poisson distribution. Obtain the MLE of the parameter of this distribution.

4. Consider a random sample $x_1, x_2, \ldots, x_n$ from a normal population having mean zero. Obtain the MLE of the variance and show that it is unbiased.

5. Consider a random sample $x_1, x_2, \ldots, x_n$ from a population following binomial distribution having parameters n and p. Find the MLE of p and show that it is unbiased.

6. Find the estimates of μ and σ in the normal populations N (μ, σ^2) by the method of moments.

7. Show that the estimates of the parameter of the Poisson distribution obtained by the method of maximum likelihood and the method of moments are identical.

8. Find a 95% C.I. for the mean of a normal population with $\sigma = 3$, given the sample 2.3, – 0.2, 0.4 and – 0.9.

9. In a sample of size 10, the sample mean is 3.22 and the sample variance 1.21. Find the 95% C.I. for the population mean.

10. A sample of size 10 from a normal population produces the data 2.03, 2.02, 2.01, 2.00, 1.99, 1.98, 1.97, 1.99, 1.96 and 1.95. From the sample find the 95% C.I. for the population mean.

11. A random sample of size 10 from a N (μ, σ^2) yields sample mean 4.8 and sample variance 8.64. Find 95% and 99% confidence intervals for the population mean.

12. The following random sample was obtained from a normal population : 12, 9, 10, 14, 11, 8. Find the 95% C.I. for the population S.D. when the population mean is (*i*) known to be 13, (*ii*) unknown.

13. The marks obtained by 15 students in an examination have a mean 60 and variance 30. Find 99% confidence interval for the mean of the population of marks, assuming it to be normal.

14. 228 out of 400 voters picked at random from a large electorate said that they were going to vote for a particular candidate. Find 95% C.I. for the proportion of voters of the electorate who would in favour of the candidate.

15. In a random sample of 300 road accidents, it was found that 114 were due to bad weather. Construct a 99% confidence interval for the corresponding true proportions.

16. A study shows that 102 of 190 persons who saw an advertisement on a product on T.V. during a sports program and 75 of 190 other persons who saw it advertised on a variety show purchased the product. Construct a 99% confidence interval for the difference of sample proportions.

Answers

1. $\hat{\alpha} = 0.890091$ **2.** $\hat{a} = \dfrac{n}{\ln\left(2^n / \prod_{i=1}^{n} x_i\right)} - 1$ **3.** $\hat{\lambda} = \bar{x}$

4. $\sigma^2 = \Sigma x_i^2 / n$ **5.** $\hat{p} = \bar{x} / n.$ **6.** $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = s^2$

NOTES

8. [–2.54, 3.34] **9.** [2.39, 4.05] **10.** [1.972, 2.008]

11. 95% C.I. [2.233, 7.367], 99% C.I. [1.616, 7.984]

12. (*i*) [1.97, 6.72], (*ii*) [1.35, 5.30] **13.** [55.64, 64.36]

14. [0.52, 0.62] **15.** [0.31, 0.45] **16.** [0.02, 0.28].

FURTHER READINGS

1. ATB of Quantitative Techniques, N.P. Bali, University Science Press.
2. Statistics and Operation Research — A Unified Approach, Dr. Debashis Dutta, Laxmi Publication.
3. Business Mathematics & Statistics, B.M. Agarwal, Ane Books Pvt. Ltd.
4. ATB of Quantitative Techniques, N.P. Bali, University Science Press.
5. Statistics and Operation Research — A Unified Approach, Dr. Debashis Dutta, Laxmi Publication.
6. Business Mathematics & Statistics, B.M. Agarwal, Ane Books Pvt. Ltd.

5. HYPOTHESIS TESTING

NOTES

STRUCTURE

Introduction
Null Hypothesis and Alternative Hypothesis
Level of Significance and Confidence Limits
Type I Error and Type II Error
Power of the Test
Test of Significance for Small Samples
Student's *t*-Test
Assumptions for Student's *t*-test
Degree of Freedom
Test for Single Mean
t-test for Difference of Means
Paired *t*-test For Difference of Means
F-test
Properties of F-distribution
Procedure to F-test
Critical Values of F-distribution
Test of Significance for Large Samples
Test of Significance for Proportion
Test of Significance for Single Mean
Test of Significance for Difference of Means

5.1. INTRODUCTION

To describe a set of data or observations, we use statistics such as mean and standard deviation. These statistics are estimated from samples. Sample is nothing but a small section selected from the population and the process of drawing or selecting a sample from the population is called 'sampling'. It is essential that a sample must be a random selection so that each member of the population has the equal chance of being selection in the sample. A statistical population consists of observations of some characteristic of interest associated with the individuals concerned and not the individual items or persons themselves.

A statistical measure based only on all the units selected in a sample is called 'statistic', *e.g.*, sample mean, sample standard deviation, proportion of defectives, etc. whereas a statistical measure based on all the units in the population is called 'parameter'. The terms like mean, median, mode, standard deviation are called parameters when they describe the characteristics of the population and are called statistic when they describe the characteristics of the sample.

A very important aspect of the sampling theory is the study of the tests of significance which enables us to decide on the basis of the sample results whether to

accept or reject the hypothesis. A test of significance can be used to compare the characteristics of two samples of the same type. Some of the well known tests of significance for small samples are *t*-test and F-test.

NOTES

5.2. NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

A statistical hypothesis is a statement about a population parameter. There are two types of statistical hypothesis, null hypothesis and alternative hypothesis.

The hypothesis formulated for the sake of rejecting it under the assumption that it is true, is called the null hypothesis and is denoted by H_0. Null hypothesis asserts that there is no significant difference between the sample statistic and the population parameter and whatever difference is observed that is merely due to fluctuations in sampling from the same population.

Rejecting null hypothesis implies that it is rejected in favour of some other hypothesis which is accepted. A hypothesis which is accepted when H_0 is rejected is called the alternative hypothesis and is denoted by H_1. What we intend to conclude is stated in the alternative hypothesis.

5.3. LEVEL OF SIGNIFICANCE AND CONFIDENCE LIMITS

The probability level below which we reject the hypothesis is known as the 'level of significance'. The region in which a sample value falling is rejected, is known as the 'critical region' or the 'rejection region'. We generally, take two critical regions which cover 5% and 1% areas of the normal curve.

Depending on the nature of the problem, we use a single-tail test or double-tail test to estimate the significance of a result. In a single-tail test, only the area on the right of an ordinate is taken into consideration whereas in a double-tail test, the areas of both the tails of the curve representing the sampling distribution are taken into consideration.

For example, a test for testing the mean of a population

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis $H_1 : \mu > \mu_0$ (right tailed) or $H_1 : \mu < \mu_0$ (left tailed) is a single tailed test. In the right tailed test ($H_1 : \mu > \mu_0$), the critical region lies entirely in the right tail of the sampling distribution ; while for the left tail test ($H_1 : \mu < \mu_0$), the critical region is entirely in the left tail of the sampling distribution.

A test of statistical hypothesis where the alternative hypothesis is two tailed such as :

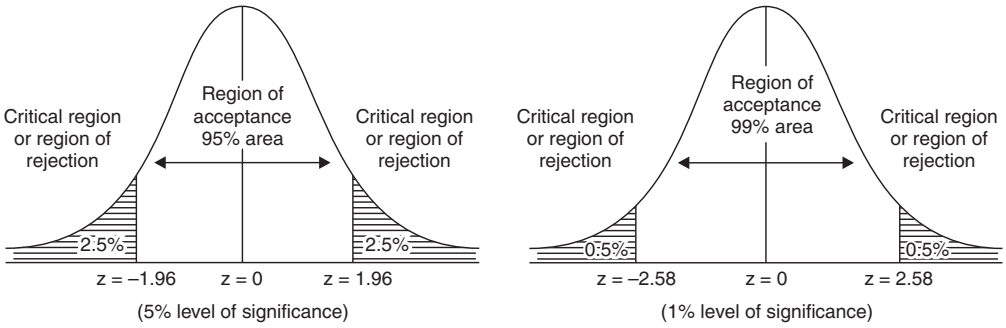
$H_0 : \mu = \mu_0$ against the alternative hypothesis

$H_1 : \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is known as two tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.

The value of z corresponding to 5% level of significance is ± 1.96 and corresponding to 1% level of significance value of z is ± 2.58. The set of z-scores outside the range

± 1.96 and ± 2.58 constitute the critical region of the hypothesis (or the region of rejection) at 5% and 1% level of significance respectively.

The following figure showing region of acceptance and rejection for 5% and 1% level of significance.



5.4. TYPE I ERROR AND TYPE II ERROR

The error of rejecting H_0 when H_0 is true is called the type I error and the error of accepting H_0 when H_0 is false (H_1 is true) is called the type II error. The probability of type I error is denoted by α and the probability of type II error is denoted by β.

P (rejecting H_0 when H_0 is true) = α

P (accepting H_0 when H_1 is true) = β

5.5. POWER OF THE TEST

A good test should accept the null hypothesis when it is true and reject the null hypothesis when it is false. $1 - \beta$ (*i.e.*, 1–probability of type II error) measures how well the test is working and is called the power of the test.

Power of the test = $1 - \beta$.

TEST OF SIGNIFICANCE FOR SMALL SAMPLES

5.6. STUDENT'S t-TEST

Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample of size n ($n < 30$) from a normal population with mean μ and variance σ^2. The student's t-test is defined as

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean and $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is an unbiased estimate of the standard deviation σ.

NOTES

5.7. ASSUMPTIONS FOR STUDENT'S t-TEST

The following assumptions are made in student's t–test :

(*i*) The parent population from which the sample is drawn is normal.

(*ii*) The population standard deviation (σ) is unknown

(*iii*) Sample size is less than 30.

5.8. DEGREE OF FREEDOM

The number of independent variates which make up the statistic is known as the degree of freedom (d.f.) and is denoted by ν (the letter 'Nu' of the Greek alphabet).

In general the degree of freedom is defined as

d.f. = number of frequencies – number of independent constraints on them.

5.9. TEST FOR SINGLE MEAN

Suppose we want to test

(*i*) If a random sample x_i ($i = 1, 2, ..., n$) of size n has been drawn from a normal population with a specified mean say μ or

(*ii*) If the sample mean differs significantly from the hypothetical value μ of the population mean.

Under null hypothesis H_0 :

(*i*) The sample mean has been drawn from the population with mean μ or

(*ii*) There is no significant difference between the sample mean $\bar{x}$ and the population mean μ, the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

follows Student's t-distribution with $(n - 1)$ degrees of freedom.

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated $|t|$ > tabulated t, H_0 is rejected and if calculated $|t|$ < tabulated t, H_0 may be accepted.

Note. We know, the sample variance

$$s^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

$$\therefore \quad ns^2 = (n-1)\,S^2$$

or

$$\frac{S^2}{n} = \frac{s^2}{n-1} \quad \Rightarrow \quad \frac{S}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

Hence, the test statistic becomes

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}.$$

NOTES

SOLVED EXAMPLES

Example 1. *The mean weekly sales of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful ?*

Solution. Here, $n = 22$, $\bar{x} = 153.7$, $s = 17.2$

Null hypothesis $H_0 : \mu = 146.3$, *i.e.*, the advertising campaign is not successful.

Alternative hypothesis $H_1 : \mu > 146.3$ (Right tail)

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \text{ with } (22 - 1) = 21 \text{ d.f.}$$

$$t = \frac{153.7 - 146.3}{17.2/\sqrt{22-1}} = \frac{7.4 \times \sqrt{21}}{17.2} = 9.$$

Since calculated value of $t = 9$ is greater than the tabulated value of $t = 1.72$ for 21 d.f. at 5% level of significance. It is highly significant. So H_0 is rejected, *i.e.*, the advertising campaign was successful in promoting sales.

Example 2. *Ten individuals are chosen at random from a normal population and the heights are found to be in inches 63, 63, 66, 67, 68, 69,70, 70, 71 and 71. Test if the sample belongs to the population whose mean height is 66 inches. (Given $t_{0.05} = 2.26$ for 9 d.f.)*

Solution.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
63	– 4.8	23.04
63	– 4.8	23.04
66	– 1.8	3.24
67	– 0.8	0.64
68	0.2	0.04
69	1.2	1.44
70	2.2	4.84
70	2.2	4.84
71	3.2	10.24
71	3.2	10.24
$\Sigma x_i = 678$		$\Sigma(x_i - \bar{x})^2 = 81.6$

Here, $n = 10$

$$\bar{x} = \text{sample mean} = \frac{\Sigma x_i}{n} = \frac{678}{10} = 67.8 \text{ inches}$$

$$S = \sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2} = \sqrt{\frac{1}{9} \times 81.6}$$

$$= \sqrt{9.0667} = 3.011$$

Null hypothesis $H_0 : \mu = 66$, *i.e.*, population mean is 66 inches

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{67.8 - 66}{3.011/\sqrt{10}} = \frac{1.8 \times \sqrt{10}}{3.011} = \frac{5.692}{3.011} = 1.8904$$

degree of freedom $= n - 1 = 10 - 1 = 9$

$t_{0.05} = 2.26$ for 9 d.f.

As the calculated value of $|t|$ is less than $t_{0.05}$, the difference between $\bar{x}$ and μ may be due to fluctuations of random sampling. H_0 may be accepted. In other words, the data does not provide any significant evidence against the hypothesis that the population mean is 66 inches.

NOTES

Example 3. *A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. (Given $t_{0.05} = 2.13$, $t_{0.01} = 2.95$ for 15 degrees of freedom)*

Solution. Here, $\bar{x} = 41.5$ inches, $n = 16$, $\Sigma(x_i - \bar{x})^2 = 135$ sq. inches

$$S = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{15} \times 135} = \sqrt{9} = 3$$

Null hypothesis $H_0 : \mu = 43.5$ inches, *i.e.*, the data are consistent with an assumption that the mean height in population is 43.5 inches.

Alternative hypothesis $H_1 : \mu \neq 43.5$ inches

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$|t| = \frac{|41.5 - 43.5|}{3/\sqrt{16}} = \frac{2 \times 4}{3} = 2.667$$

degrees of freedom $= n - 1 = 16 - 1 = 15$

We are given $t_{0.05} = 2.13$ and $t_{0.01} = 2.95$ for 15 degrees of freedom.

Since calculated $|t|$ is greater than $t_{0.05} = 2.13$, null hypothesis H_0 is rejected at 5% level of significance and we conclude that the assumption of mean 43.5 inches for the population is not reasonable.

Remark. Since calculated $|t|$ is less than $t_{0.01} = 2.95$, null hypothesis H_0 may be accepted at 1% level of significance.

5.10. t-TEST FOR DIFFERENCE OF MEANS

Given two independent random samples x_i ($i = 1, 2, \ldots\ldots, n_1$) and y_j ($j = 1, 2, \ldots\ldots, n_2$) of sizes n_1 and n_2 with means $\bar{x}$ and $\bar{y}$ and standard deviations S_1 and S_2 from normal populations with the same variance, we have to test the hypothesis that the population means are same. In other words, since a normal distribution is completely specified by its mean and variance, we have to test the hypothesis that the two independent samples come from the same normal population.

The statistic is given by

$$t = \frac{\bar{x} \sim \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \; ; \; \bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j$$

and $$S^2 = \frac{1}{(n_1 + n_2 - 2)} [(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2]$$

or $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2\right]$$

follows Student's t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

If the calculated value of $|t|$ be > tabulated t, the difference between the sample means is said to be significant at certain level of significance ; otherwise the data are said to be consistent with the hypothesis.

NOTES

5.11. PAIRED t-TEST FOR DIFFERENCE OF MEANS

If the size of the two samples is the same, say equal to n, and the data are paired, *i.e.* (x_i, y_i), $(i = 1, 2, ..., n)$ corresponds to the same ith sample unit. The problem is to test if the sample means differ significantly or not.

Here, we consider the increments, $d_i = x_i - y_i$, $(i = 1, 2, ..., n)$.

Under the null hypothesis H_0 that increments are due to fluctuations of sampling, the statistic

$$t = \frac{\bar{d}}{S/\sqrt{n}},$$

where $$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2$$

follows Student's t-distribution with $(n - 1)$ degrees of freedom. If Σd_i is negative, we may consider $|\bar{d}|$. This test is generally one tailed test. Therefore, the alternative hypothesis is $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$.

SOLVED EXAMPLES

Example 1. *The following data related to the heights (in cms) of two different varieties of wheat plants.*

Variety 1	*63*	*65*	*68*	*69*	*71*	*72*				
Variety 2	*61*	*62*	*65*	*66*	*69*	*69*	*70*	*71*	*72*	*73*

Test the null hypothesis that the mean heights of plants of both varieties are the same.

Solution. Given $n_1 = 6$, $n_2 = 10$

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 > \mu_2$ (right tail)

Under H_0 the test statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

NOTES

Variety 1

x	$x - \bar{x} = x - 68$	$(x - \bar{x})^2$
63	– 5	25
65	– 3	9
68	0	0
69	1	1
71	3	9
72	4	16
$\Sigma x = 408$		$\Sigma(x - \bar{x})^2 = 60$

Variety 2

y	$y - \bar{y} = y - 67$	$(y - \bar{y})^2$
61	– 6	36
62	– 5	25
65	– 2	4
65	– 2	4
66	– 1	1
66	– 1	1
70	3	9
70	3	9
72	5	25
73	6	36
$\Sigma y = 670$		$\Sigma(y - \bar{y})^2 = 150$

$$\bar{x} = \frac{1}{n_1}\Sigma x_i = \frac{408}{6} = 68 \qquad \bar{y} = \frac{1}{n_2}\Sigma y_i = \frac{670}{10} = 67$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\,[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2]$$

$$= \frac{1}{6 + 10 - 2}\,[60 + 150] = \frac{210}{14} = 15 \quad \Rightarrow \quad S = 3.873$$

$$\therefore \qquad t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{68 - 67}{3.873\sqrt{\frac{1}{6} + \frac{1}{10}}} = \frac{1}{3.873 \times 0.5164} = 0.499$$

Tabulated $t_{0.05}$ for 14 degrees of freedom for single tail-test is 1.76.

Since calculated value of t is less than 1.76, it is not at all significant at 5% level of significance. Hence, H_0 may be accepted and we conclude that the height of the plants are not different at 5% level of significance.

Example 2. *The mean values of birth weight with standard deviations and sample sizes are given below by socio-economic status. Is the mean difference in birth weight significant between socio-economic group ?*

	High socio-economic graph	*Low socio-economic group*
Sample size	$n_1 = 15$	$n_2 = 10$
Birth weight (kg)	$\bar{x} = 2.91$	$\bar{y} = 2.26$
Standard deviation	$S_1 = 0.27$	$S_2 = 0.22$

Solution. Given $n_1 = 15$, $n_2 = 10$, $\bar{x} = 2.91$, $\bar{y} = 2.26$

$S_1 = 0.27$ and $S_2 = 0.22$

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 > \mu_2$ (right tail), *i.e.* high socio-economic group is superior to low socio-economic group.

Under H_0 the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

NOTES

$$S^2 = \frac{1}{n_1 + n_2 - 2}\,[(n_1 - 1)\,S_1^{\,2} + (n_2 - 1)\,S_2^{\,2}]$$

$$= \frac{1}{15 + 10 - 2}\,[(15 - 1) \times (0.27)^2 + (10 - 1) \times (0.22)^2]$$

$$= \frac{1.0206 + 0.4356}{23} = \frac{1.4562}{23} = 0.063$$

$$\Rightarrow \qquad S = 0.25$$

$$\therefore \qquad t = \frac{2.91 - 2.26}{0.25\sqrt{\frac{1}{15} + \frac{1}{10}}} = \frac{0.65 \times \sqrt{150}}{0.25 \times \sqrt{25}} = \frac{0.65 \times 2.45}{0.25} = 6.37$$

Tabulated value of t for 23 degrees of freedom at 5% level of significance for right tailed test is 1.71. Since calculated t is much greater than tabulated t, it is highly significant and H_0 is rejected and conclude that mean of high group is greater than low group.

Example 3. *In a test examination given to two groups of students, the marks obtained were as follows :*

Group I	*25*	*32*	*30*	*34*	*24*	*14*	*32*	*24*	*30*	*31*	*35*	*25*			
Group II	*44*	*34*	*22*	*10*	*47*	*31*	*40*	*30*	*32*	*35*	*18*	*21*	*35*	*29*	*22*

Examine the significance of difference between the arithmetic average of marks secured by students of the two groups.

Solution. Here, $n_1 = 12$, $n_2 = 15$

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two-tailed)

x	$x - \bar{x} = x - 28$	$(x - \bar{x})^2$
25	– 3	9
32	4	16
30	2	4
34	6	36
24	– 4	16
14	– 14	196
32	4	16
24	– 4	16
30	2	4
31	3	9
35	7	49
25	– 3	9
$\Sigma x = 336$	$\Sigma(x - \bar{x}) = 0$	$\Sigma(x - \bar{x})^2 = 380$

y	$y - \bar{y} = y - 30$	$(y - \bar{y})^2$
44	14	196
34	4	16
22	– 8	64
10	–20	400
47	17	289
31	1	1
40	10	100
30	0	0
32	2	4
35	5	25
18	– 12	144
21	– 9	81
35	5	25
29	– 1	1
22	– 8	64
$\Sigma y = 450$	$\Sigma(y - \bar{y}) = 0$	$\Sigma(y - \bar{y})^2 = 1410$

NOTES

$$\bar{x} = \frac{1}{n_1}\Sigma x_i = \frac{336}{12} = 28, \quad \bar{y} = \frac{1}{n_2}\Sigma y_i = \frac{450}{15} = 30$$

Under H_0, the test statistic is

$$t = \frac{\bar{x} \sim \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right]$$

$$= \frac{1}{12 + 15 - 2}[380 + 1410] = \frac{1790}{25} = 71.6$$

$$\Rightarrow \quad S = 8.46$$

$$\therefore \quad t = \frac{30 - 28}{8.46\sqrt{\frac{1}{12} + \frac{1}{15}}} = \frac{2}{8.46 \times 0.387} = 0.61$$

Tabulated value of $t_{0.05}$ for 25 degrees of freedom is 2.06.

Since calculated value of t is less than tabulated value of t at 5% level of significance. H_0 may be accepted and we may conclude that two averages do not differ significantly.

Example 4. *Memory capacity of 8 students was tested before and after training. State at 5% level of significance whether the training was effective from the following scores :*

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
Before	*49*	*53*	*51*	*52*	*47*	*50*	*52*	*53*	*407*
After	*52*	*55*	*52*	*53*	*50*	*54*	*54*	*53*	*423*

Use paired t-test for your answer.

Solution. Let x denotes the scores before training and y denotes the scores after training.

Null hypothesis $H_0 : \mu_1 = \mu_2$, *i.e.* there is no significant difference in the scores before and after the training. In other words, the given increments are just by chance (fluctuations of sampling).

Alternative hypothesis $H_1 : \mu_1 < \mu_2$ (to conclude that training has been effected) (One tail)

Student	*Score before training (x)*	*Score after training (y)*	$d = x - y$	d^2
1	49	52	– 3	9
2	53	55	– 2	4
3	51	52	– 1	1
4	52	53	– 1	1
5	47	50	– 3	9
6	50	54	– 4	16
7	52	54	– 2	4
8	53	53	0	0
			$\Sigma d = -16$	$\Sigma d^2 = 44$

Under H_0 the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

NOTES

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{-16}{8} = -2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}[\Sigma d_i^2 - n(\bar{d})^2]$$

$$= \frac{1}{7}[44 - 8 \times (-2)^2] = \frac{44-32}{7} = \frac{12}{7} = 1.714$$

$\Rightarrow$ $S = 1.31$

$\therefore$ $$|t| = \frac{|d|}{S/\sqrt{n}} = \frac{|-2|}{1.31/\sqrt{8}} = \frac{2 \times 2.83}{1.31} = 4.32$$

Tabulated $t_{0.05}$ for $(8 - 1) = 7$ degrees of freedom for one tail test is 1.90.

Since calculated value of t is greater than the tabulated t, H_0 is rejected at 5% level of significance. Hence, we conclude that the scores differ significantly before and after the training, *i.e.* training was effected.

Example 5. *A certain drug administred to 10 patients showed the following additional hours of sleep :*

– 1.0, 0.5, 2.7, – 0.6, 1.2, 1.8, 1.6, 3.5, 0.2, – 1.7

Can it be concluded that the drug does produce additional hours of sleep ?

Solution. Here, d_i are given as

$$d_i = x_i - y_i = -1.0, 0.5, 2.7, -0.6, 1.2, 1.8, 1.6, 3.5, 0.2, -1.7$$

$$n = 10$$

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{-1.0 + 0.5 + 2.7 - 0.6 + 1.2 + 1.8 + 1.6 + 3.5 + 0.2 - 1.7}{10}$$

$$= \frac{8.2}{10} = 0.82$$

$\Sigma d_i^2 = 1 + 0.25 + 7.29 + 0.36 + 1.44 + 3.24 + 2.56 + 12.25 + 0.04 + 2.89 = 31.32$

Null hypothesis $H_0 : \mu_1 = \mu_2$, *i.e.* the drug does not produce any additional hours of sleep.

Alternative hypothesis $H_1 : \mu_1 < \mu_2$, *i.e.* drug is effective (one tail).

Under H_0, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}[\Sigma d_i^2 - n(\bar{d})^2]$$

$$= \frac{1}{10-1}[31.32 - 10 \times (0.82)^2] = \frac{1}{9}[31.32 - 6.724] = 2.733$$

$\Rightarrow$ $S = 1.653$

$\therefore$ $$t = \frac{0.82 \times \sqrt{10}}{1.653} = \frac{2.593}{1.653} = 1.57$$

NOTES

Tabulated $t_{0.05} = 1.833$ with $(10 - 1)$ degrees of freedom at 5% level of significance.

Since calculated value of t is less than the tabulated t, H_0 is accepted at 5% level of significance. Hence, we conclude that the drug do not produce additional hours of sleep.

EXERCISE 5.1

1. A brand of matches is sold in boxes on which it is claimed that the average contents are 40 matches. A check on a pack of 5 boxes gives the following results :

41, 39, 37, 40, 38

(*i*) Test the manufacturer's claim keeping the interests of both the manufacturer and the customer in mind.

(*ii*) As a customer test the manufacturer's claim.

2. A sample of size 10 drawn from a normal population has a mean 31 and a variance 2.25. Is it reasonable to assume that the mean of the population is 30 ? (Use 1% level of significance).

3. A random sample of size 10 from a normal population with mean μ gives a sample mean of 40 and sample standard deviation of 6. Test the hypothesis that $\mu = 44$ against $\mu \neq 44$ at 5% level of significance.

4. A new drug manufacturer wants to market a new drug only if he could be quite sure that the mean temperature of a healthy person taking the drug could not rise above 98.6°F otherwise he will withhold the drug. The drug is administered to a random sample of 17 healthy persons. The mean temperature was found to be 98.4°F with a standard deviation of 0.6°F. Assuming that the distribution of the temperature is normal and $\alpha = 0.01$, what should the manufacturer do ?

5. The marks of students in two groups were obtained as

I	18	20	36	50	49	36	34	49	41
II	29	28	26	35	30	44	46		

Test whether the groups were identical.

(Given $t_{0.05} = 2.14$ for 14 degrees of freedom)

6. Two different types of drugs A and B were tried on certain patients for increasing weight. 5 persons were given drug A and 7 persons were given drug B. The increase in weight in pounds is given below :

Drug A	8	12	13	9	3		
Drug B	10	8	12	15	6	8	11

Do the two drugs differ significantly with regard to their effect in increasing weight.

(Given $t_{0.05} = 2.23$ for 10 degrees of freedom)

7. The mean life of a sample of 10 electric light bulbs was found to be 1456 hours with standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with standard deviation of 398 hours. Is there a significant difference between the means of the two batches ?

(Given $t_{0.05} = 2.06$ for 25 degrees of freedom)

8. To verify whether a course in Statistics improved performance, a similar test was given to 12 participants both before and after the course. The original marks recorded in alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. After the course, the marks were in the same order 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. Was the course useful ?

(Given $t_{0.05} = 2.201$ for 11 degrees of freedom)

9. A certain medicine given to each of the 9 patients resulted in the following increase of blood pressure. Can it be concluded that the medicine will in general be accompanied by an increase in blood pressure.

7, 3, – 1, 4, – 3, 5, 6, – 4, – 1

(Given $t_{0.05} = 2.306$ for 8 degrees of freedom)

NOTES

Answers

1. (*i*) Accept manufacturer 's claim (*ii*) manufacturer's claim is justified.

2. Yes **3.** Accept null hypothesis

4. The manufacturer should market the drug **5.** Two groups are identical

6. No **7.** No **8.** Yes **9.** No

5.12. F-TEST

This test uses the variance ratio to test the significance of difference between two sampled variances. F-test which is based on F-distribution is called so in honour of a great statistician Prof. R.A. Fisher.

Let $x_1, x_2, \ldots\ldots, x_{n_1}$ and $y_1, y_2, \ldots\ldots, y_{n_2}$ be the values of two independent random samples drawn from the same normal population with variance σ^2. Then, we define variance ratio F as follows :

$$F = \frac{S_1^2}{S_2^2} \; ; S_1 > S_2,$$

where

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

and $\bar{x}, \bar{y}$ are the sample means.

The distribution of variance ratio F with ν_1 and ν_2 degrees of freedom is given by

$$y = \frac{y_0 \, F^{\left(\frac{\nu_1 - 2}{2}\right)}}{\left(1 + \frac{\nu_1}{\nu_2} F\right)^{\left(\frac{\nu_1 + \nu_2}{2}\right)}},$$

where y_0 is so chosen that the total area under the curve is unity.

The parameters ν_1 and ν_2 represent degrees of freedom. For samples of sizes n_1 and n_2, we have

$$\nu_1 = n_1 - 1 \quad \text{and} \quad \nu_2 = n_2 - 1.$$

5.13. PROPERTIES OF F-DISTRIBUTION

(*i*) The value of F cannot be negative as both terms of F-ratio are the squared values.

(*ii*) The range of the values of F is from 0 to ∞.

(*iii*) The F-distribution is independent of the population variance σ^2 and depends on ν_1 and ν_2 only.

NOTES

The F-distribution for various degrees of freedom ν_1 and ν_2 is given in the following table :

Table : Values of F for 5% and 1% level, where ν_1 is the number of degree of freedom for greater estimate of variance and ν_2 for the smaller estimate of variance.

5.14. PROCEDURE TO F-TEST

(*i*) Set up the null hypothesis $H_0 = \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* the independent estimates of the common population variance do not differ significantly.

(*ii*) Find the degrees of freedom ν_1 and ν_2 given by $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ respectively.

(*iii*) Calculate the variances of two samples and then calculate F.

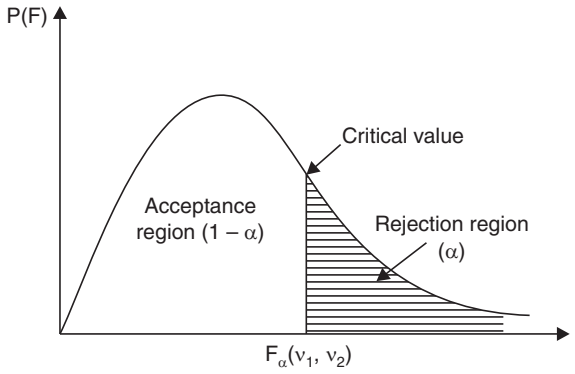
(*iv*) From F-distribution table note the value of F for ν_1, ν_2 degrees of freedom at the desired level of significance.

(*v*) Compare the calculated value of F with tabulated value of F at the desired level of significance. If the calculated value of F is less than the tabulated value, then the difference is not significant and we may conclude that the same could have come from two populations with the same variance *i.e.*, accept H_0, otherwise reject H_0.

5.15. CRITICAL VALUES OF F-DISTRIBUTION

The available F-table give the critical values of F for the right-tailed test, *i.e.* the critical region is determined by the right-tail areas. Thus, the significance value $F_\alpha(\nu_1, \nu_2)$ at level of significance and (ν_1, ν_2) degrees of freedom is determined by

$P[F > F_\alpha(\nu_1, \nu_2)] = \alpha$, as shown below :



SOLVED EXAMPLES

Example 1. *In one sample of size 8 the sum of the squares of deviations of the sample values from the sample mean is 84.4 and in the other sample of size 10 it is 102.6. Test whether this difference is significance at 5% level. Given that for* $\nu_1 = 7$ *and* $\nu_2 = 9$ *;* $F_{0.05} = 3.29$.

Solution. Here, $n_1 = 8$, $n_2 = 10$

and $\Sigma(x - \bar{x})^2 = 84.4$, $\Sigma(y - \bar{y})^2 = 102.6$

NOTES

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{7} \times 84.4 = 12.057$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{9} \times 102.6 = 11.4$$

Under $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* the estimates of σ^2 given by the samples are homogeneous,

$$F = \frac{S_1^2}{S_2^2} = \frac{12.057}{11.4} = 1.057$$

For $\nu_1 = 7$ and $\nu_2 = 9$, we have $F_{0.05} = 3.29$. Since calculated value of F is less than $F_{0.05}$, H_0 may be accepted at 5% level of significance.

Example 2. *Two random samples gave the following information :*

Sample	*Size*	*Sample mean*	*Sum of squares of deviations from the mean*
1	*10*	*15*	*90*
2	*12*	*14*	*108*

Test whether the samples have been drawn from the same normal population. Given that for $\nu_1 = 9$ and $\nu_2 = 11$; $F_{0.05} = 2.90$ (approx.).

Solution. Here, $n_1 = 10$, $n_2 = 12$, $\bar{x} = 15$, $\bar{y} = 14$

$$\Sigma (x - \bar{x})^2 = 90 \; ; \Sigma(y - \bar{y})^2 = 108$$

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{9} \times 90 = 10$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{11} \times 108 = 9.82$$

Under $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* two samples have been drawn from the same normal population.

$$F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018$$

For $\nu_1 = 9$ and $\nu_2 = 11$, we have $F_{0.05} = 2.90$.

Since calculated value of F is less than $F_{0.05}$ it is not significant. Hence, null hypothesis H_0 may be accepted.

Example 3. *The samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 and 91 square units respectively. Test whether the samples have been drawn from the same normal population. Given that for $\nu_1 = 8$ and $\nu_2 = 7$; $F_{0.05} = 3.73$.*

Solution. Here, $n_1 = 9$, n_2, $= 8$, $\Sigma(x - \bar{x})^2 = 160$, $\Sigma(y - \bar{y})^2 = 91$

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{1}{8} \times 160 = 20$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1}{7} \times 91 = 13$$

NOTES

Under $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, *i.e.* two samples have been drawn from the same normal population.

$$F = \frac{S_1^2}{S_2^2} = \frac{20}{13} = 1.54 \text{ (approx.)}$$

For $\nu_1 = 8$ and $\nu_2 = 7$, we have $F_{0.05} = 3.73$

Since calculated value of F is less than $F_{0.05}$ it is not significant. Hence, H_0 may be accepted.

Example 4. *Two samples are drawn from two normal populations. From the following data test whether the two samples have the same variances at 5% level of significance.*

Sample I	*60*	*65*	*71*	*74*	*76*	*82*	*85*	*87*		
Sample II	*61*	*66*	*67*	*85*	*78*	*88*	*86*	*85*	*63*	*91*

Solution. Here, $n_1 = 8$, $n_2 = 10$

Under $H_0 : S_1^2 = S_2^2$, *i.e.* two samples have the same variance.

$$H_1 : S_1^2 \neq S_2^2$$

Sample-I			Sample-II		
x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$
60	60–75 = – 15	225	61	61–77 = – 16	256
65	65–75 = – 10	100	66	66–77 = – 11	121
71	71–75 = – 4	16	67	67–77 = – 10	100
74	74–75 = – 1	1	85	85–77 = 8	64
76	76–75 = 1	1	78	78–77 = 1	1
82	82–75 = 7	49	88	88–77 = 11	121
85	85–75 = 10	100	86	86–77 = 9	81
87	87–75 = 12	144	85	85–77 = 8	64
			63	63–77 = – 14	196
			91	91–77 = 14	196
$\Sigma x = 600$		$\Sigma(x - \bar{x})^2 = 636$	$\Sigma y = 770$		$\Sigma(y - \bar{y})^2 = 1200$

$$\bar{x} = \frac{\Sigma x}{n_1} = \frac{600}{8} = 75 \qquad \bar{y} = \frac{\Sigma y}{n_2} = \frac{770}{10} = 77$$

$$\text{Variance of sample-I} = S_1^2 = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{636}{8 - 1} = 90.857$$

$$\text{Variance of sample-II} = S_2^2 = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{1200}{10 - 1} = 133.33$$

$$F = \frac{S_2^2}{S_1^2} = \frac{133.33}{90.857} = 1.467$$

For $\nu_1 = 7$ and $\nu_2 = 9$, we have $F_{0.05} = 3.29$.

Since calculated value of F is less than $F_{0.05}$, H_0 may be accepted, *i.e.* the samples I and II have the same variance.

NOTES

EXERCISE 5.2

1. In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference is significant at 5% level.

2. The following are the values in thousands of an inch obtained by two engineers in 10 successive measurements with the same micrometer. Is one engineer significantly more consistent than the other ?

Engineer A	503	505	497	505	495	502	499	493	510	501
Engineer B	502	497	492	498	499	495	497	496	498	

3. The nicotine content (in milligrams) of two samples of tobacco were found to be as follows :

Sample A	24	27	26	21	25	
Sample B	27	30	28	31	22	36

Can it be said that the two samples come from the same normal population ?

4. The daily wages in ₹ of skilled workers in two cities are as follows :

City	*Size of sample of workers*	*S.D. of wages in the sample*
A	16	25
B	13	32

Test at 5% level of significance the equality of variances of the wage distribution in the two cities.

5. The time taken by workers in performing a job by methods I and II is given below :

Method I	20	16	26	27	23	22	–
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population from which these samples are drawn do not differ significantly ?

6. Two random samples drawn from two normal populations are given below :

Sample I	63	65	68	69	71	72	–	–	–	–
Sample II	63	62	65	66	69	69	70	71	72	73

Test whether the two populations have the same variance at 5% level of significance.

Answers

1. No **2.** Not significant **3.** yes
4. Accepted **5.** Not significant **6.** Yes.

NOTES

TEST OF SIGNIFICANCE FOR LARGE SAMPLES

For practical purposes a sample is taken as a large sample if $n > 30$. Under large sample test there are some important tests to test the significance. These tests are as follows :

1. Test of significance for proportion
   (*i*) Single proportion (*ii*) Difference of proportions
2. Test of significance for single mean.
3. Test of significance for differences of
   (*i*) Means (*ii*) Standard deviations.

5.16. TEST OF SIGNIFICANCE FOR PROPORTION

(*i*) Single proportion

This test is used to test the significant difference between proportion of the sample and the population.

Let X be the number of successes in n independent trials with constant probability P of success for each trial.

We have $E(X) = nP$ and $V(X) = nPQ$, where $Q = 1 - P$ = probability of failure

Now, $$p = \frac{X}{n}$$ (p = observed proportion of success)

Now, $$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{nP}{n} = P$$

$$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{nPQ}{n^2} = \frac{PQ}{n}$$

$$\text{S.E. } (p) = \sqrt{\frac{PQ}{n}}$$

$$Z = \frac{p - E(p)}{\text{S.E. } (p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0, 1)$$

where E → expected value, V → Variance and S.E. → Standard error

Z is called a test statistic which is used to test the significant difference of the sample and population proportion.

Note 1. The probable limits for the observed proportion of success are $E(p) \pm Z_\alpha \sqrt{V(p)}$ *i.e.*, $P \pm Z_\alpha \sqrt{\frac{PQ}{n}}$, where Z_α is the significant value at the level of significance α.

2. If P is not known then the probable limits for the proportion in the population are $p \pm Z_\alpha \sqrt{\frac{pq}{n}}$.

3. If α is not given, then we can use 3σ limits. Hence, probable limits for the observed proportion of success are $P \pm 3\sqrt{\frac{PQ}{n}}$ and probable limits for the proportion in the population are $p \pm 3\sqrt{\frac{pq}{n}}$.

4. A set of four selected values is commonly used for α. Each α and corresponding Z_α and $Z_{\alpha/2}$ values are given in the following table :

For two-tailed test		*For one-tailed test*	
α	$Z_{\alpha/2}$	α	Z_α
0.20	1.282	0.10	1.282
0.10	1.645	0.05	1.645
0.05	1.960	0.025	1.960
0.01	2.576	0.01	2.326

NOTES

(*ii*) Difference of Proportions

This test is used to test the difference between the sample proportions.

Let two samples X_1 and X_2 of sizes n_1 and n_2 respectively taken from two different populations, then $p_1 = \frac{X_1}{n_1}$ and $p_2 = \frac{X_2}{n_2}$.

To test the significance of the difference between the sample proportions p_1 and p_2 we set the null hypothesis H_0, that there is no significant difference between the two sample proportion.

Under the null hypothesis H_0, the test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{and} \quad Q = 1 - P$$

If sample proportions are not given, we set the null hypothesis

$$H_0 : p_1 = p_1$$

under H_0 the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}, \text{ where } Q_1 = 1 - P_1 \text{ and } Q_2 = 1 - P_2.$$

SOLVED EXAMPLES

Example 1. *A coin is tossed 324 times and the head turned up 175 times. Test the hypothesis that the coin is unbiased.*

Solution. Null hypothesis H_0 : the coin is unbiased *i.e.,*

$$P = \frac{1}{2}$$

Here, n = 324, X = Number of heads = 175

$$P = \text{prob. of getting a head in a toss} = \frac{1}{2}$$

$$Q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2}$$

NOTES

$$\therefore \quad Z = \frac{X - E(X)}{SE \text{ of } X} = \frac{X - nP}{\sqrt{nPQ}} = \frac{175 - 324 \times \frac{1}{2}}{\sqrt{324 \times \frac{1}{2} \times \frac{1}{2}}}$$

$$= \frac{13}{9} = 1.44 < 1.96$$

Since | Z | < 1.96, null hypothesis is accepted at 5% level of significance. Hence the coin is unbiased.

Example 2. *A die is thrown 1000 times and a throw of 5 or 6 was obtained 420 times. On the assumption of random throwing do the data indicate an unbiased die ?*

Solution. Null hypothesis H_0 : the die is unbiased

Under H_0, P = probability of getting 5 or 6

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

$$Q = 1 - P = 1 - \frac{1}{3} = \frac{2}{3}$$

Here, n = 1000, X = Number of success = 420

$$Z = \frac{X - nP}{\sqrt{nPQ}} = \frac{420 - 1000 \times \frac{1}{3}}{\sqrt{1000 \times \frac{1}{3} \times \frac{2}{3}}} = \frac{420 - 333.33}{\sqrt{222.222}} = \frac{86.67}{14.91} = 5.813$$

Since | Z | = 5.813 > 3 (Maximum value of Z), H_0 is rejected *i.e.*, the die is biased.

Example 3. *A manufacturer claims that only 4% of his products supplied by him are defective. A random sample of 600 products contained 36 defectives. Test the claim of manufacturer.*

Solution. Here p = sample proportion of defectives = $\frac{36}{600} = 0.06$

P = proportion of defectives in the population = $\frac{4}{100} = 0.04$

Q = 1 – P = 1 – 0.04 = 0.96

n = 600

Null hypothesis H_0 : P = 0.04 is true *i.e.*, the claim of manufacturer is right

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.06 - 0.04}{\sqrt{\frac{0.04 \times 0.96}{600}}} = \frac{0.02}{0.008} = 2.5$$

If we set the alternative hypothesis H_1 : P ≠ 0.04 we apply two tailed test.

Since | Z | = 2.5 > 1.96, H_0 is rejected at 5% level of significance *i.e.*, manufacturer's claim is not acceptable.

If we set the alternative hypothesis H_1 : P > 0.04 we apply right tailed test.

| Z | = 2.5 > 1.645, H_0 is rejected at 5% level of significance. *i.e.*, manufacturer's claim is not acceptable.

Example 4. *500 apples are taken at random from a large basket and 65 are found to be bad. Find the S.E. of the proportion of bad ones in a sample of this size and assign limits within which the percentage of bad apples most probably lies.*

Solution. Here, $n = 500$, X = number of bad apples in the sample = 65

p = proportion of bad apples in the sample = $\frac{65}{500} = 0.13$ and

$q = 1 - p = 1 - 0.13 = 0.87$

$\because$ The proportion of bad apples P in the population is not known.

$\therefore$ We can take P = $p = 0.13$, Q = $q = 0.87$ and N = $n = 500$

S.E. of proportion $= \sqrt{\frac{PQ}{N}} = \sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$

Limits for proportions of bad apples in the population is

$$P \pm 3\sqrt{\frac{PQ}{N}} = 0.13 \pm 3\sqrt{\frac{0.13 \times 0.87}{500}} = 0.13 \pm 0.045 = 0.175 \text{ and } 0.085$$

$= 17.5\%$ and 8.5%.

Example 5. *A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample of 300 equipments revealed that 27 are faulty. Test his claim at a significance level of (i) 5% (ii) 1%.*

Solution. Here,

X = number of equipments conforming to specifications in the samples

$= 300 - 27 = 273$

p = sample proportion conforming to specifications = $\frac{273}{300} = 0.91$

Null hypothesis H_0 : P = 0.95 (the proportion of equipments conforming to specification in the population is 95%)

$Q = 1 - P = 0.05$

H_1 : $P < 0.95$ (at least 95% conformed to specification)

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.91 - 0.95}{\sqrt{\frac{0.95 \times 0.05}{300}}} = \frac{-0.04}{0.0126} = -3.175$$

$| Z | = | -3.175 | = 3.175$

(*i*) Since the H_1 is one tailed and the significant value of Z at 5% level of significance for one tail is 1.645.

Now $| Z | = 3.175 > 1.645$, H_0 is rejected *i.e.*, manufacturer's claim is not acceptable.

(*ii*) The significant value of Z at 1% level of significance for one tail is 2.33.

Now $| Z | = 3.175 > 2.33$, H_0 is rejected *i.e.*, manufacturer's claim is not acceptable.

Example 6. *Before an increase in excise duty on tea, 400 people out of a sample of 500 persons where found to be tea drinkers. After an increase in the excise duty, 400 persons were known to be tea drinkers in a sample of 600 people. Do you think that there has been a significant decrease in the consumption of tea after the increase in the excise duty ?*

Solution. Here $n_1 = 500$, $n_2 = 600$

$X_1 = 400$, $X_2 = 400$

NOTES

p_1 = proportion of drinkers in first sample $= \frac{400}{500} = \frac{4}{5} = 0.8$

p_2 = proportion of drinkers in second sample $= \frac{400}{600} = \frac{2}{3} = 0.67$

Since proportion P of the population is not given, it can be estimated by using

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 + 400}{500 + 600} = \frac{800}{1100} = \frac{8}{11}$$

and $$Q = 1 - P = 1 - \frac{8}{11} = \frac{3}{11}$$

Null hypothesis $H_0 : P_1 = P_2$ (there is no significant difference in the consumption of tea before and after increase of excise duty)

Alternative hypothesis $H_1 : P_1 > P_2$ (right tailed test), under H_0 the test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8 - 0.67}{\sqrt{\frac{8}{11} \times \frac{3}{11}\left(\frac{1}{500} + \frac{1}{600}\right)}} = \frac{0.13}{0.027} = 4.815$$

Since | Z | = 4.815 > 1.645 also | Z | = 4.815 > 2.33 at both the significant values of Z at 5% and 1% level of significant respectively, H_0 is rejected *i.e.*, there is a significant decrease in the consumption of tea due to increase in excise duty.

Example 7. *During a country wide investigation the incidence of a chronic decease was found to be 1%. In a village of 400 strength 5 were reported to be affected where-as in another village of 1200 strength 10 were reported to be affected. Does this indicate any significant difference.*

Solution. Here, $P = 0.01$ and $Q = 1 - P = 1 - 0.01 = 0.99$

$$n_1 = 400, p_1 = \frac{5}{400} = 0.0125$$

$$n_2 = 1200, p_2 = \frac{10}{1200} = 0.0083$$

Null hypothesis $H_0 : P_1 = P_2$ (there is no significant difference)

Alternative hypothesis $H_1 : P_1 \neq P_2$ (two tailed test)

Under H_0 the test statistic $$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.0125 - 0.0083}{\sqrt{0.01 \times 0.99\left(\frac{1}{400} + \frac{1}{1200}\right)}}$$

$$= \frac{0.0042}{0.00574} = 0.732$$

Since | Z | = 0.732 < 1.96, null hypothesis is accepted at 5% level of significance. Hence the difference is not significant.

Example 8. *500 articles from a factory are examined and found to be 2% defective. 800 similar articles from a second factory are found to have only 1.5% defectives. Can it reasonably concluded that the products of the first factory are inferior to those of second ?*

NOTES

Solution. Here, $n_1 = 500$,

p_1 = proportion of defectives from first factory = $\frac{2}{100} = 0.02$

$n_2 = 800$,

p_2 = proportion of defectives from second factory = $\frac{15}{100} = 0.015$

Since proportion P of the population is not given it can be estimated by using

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{10 + 12}{500 + 800} = \frac{22}{1300} = 0.017$$

and $$Q = 1 - P = 1 - 0.017 = 0.983$$

Null hypothesis $H_0 : P_1 = P_2$ (there is no significant difference between the products of first and second factory)

Alternative hypothesis $H_1 : P_1 \neq P_2$ (two tailed test)

Under H_0 the test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.02 - 0.015}{\sqrt{0.017 \times 0.983\left(\frac{1}{500} + \frac{1}{800}\right)}}$$

$$= \frac{0.005}{0.00737} = 0.678$$

Since $|Z| = 0.678 < 1.96$, null hypothesis is accepted at 5% level of significance. Hence there is no significant difference between the products of first and second factory *i.e.*, the products of the first factory are not inferior to those of second.

Example 9. *A manufacturing firm claims that its brand. A products outsells its brand B products by 8%. If it is found that 84 out of a sample of 400 persons prefer brand A and 36 out of another sample of 200 persons prefer brand B. Test whether the 8% difference is a valid claim.*

Solution. Here, $n_1 = 400$, $n_2 = 200$

p_1 = proportion of preference of brand A = $\frac{84}{400} = 0.21$

p_2 = proportion of preference of brand B = $\frac{36}{200} = 0.18$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{84 + 36}{400 + 200} = \frac{120}{600} = 0.2$$

and $$Q = 1 - P = 1 - 0.2 = 0.8$$

Null hypothesis H_0 : 8% difference is there in the sales of brand A and brand B *i.e.*, $P_1 - P_2 = 0.08$

Alternative hypothesis $H_1 : P_1 - P_2 \neq 0.08$ (two tailed test)

NOTES

Under H_0 the test statistic

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.21 - 0.18) - (0.08)}{\sqrt{0.2 \times 0.8\left(\frac{1}{400} + \frac{1}{200}\right)}}$$

$$= -\frac{0.05}{0.0346} = -1.44$$

Since $|Z| = 1.44 < 1.96$, null hypothesis is accepted at 5% level of significance. Hence the claim of 8% difference in the sales of brand A and brand B is valid.

Example 10. *In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1400 and 1000 respectively from the two populations.*

Sol. Here, $n_1 = 1400$, $n_2 = 1000$

P_1 = proportion of fair haired in the first population = $\frac{30}{100} = 0.3$

P_2 = proportion of fair haired in the second population = $\frac{25}{100} = 0.25$

$Q_1 = 1 - P_1 = 1 - 0.3 = 0.7$, $Q_2 = 1 - P_2 = 1 - 0.25 = 0.75$

Null hypothesis $H_0 : p_1 = p_2$ (Sample proportions are equal) *i.e.*, the difference in population proportions is likely to be hidden in sampling.

Alternative hypothesis $H_1 : p_1 \neq p_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} = \frac{0.30 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1400} + \frac{0.25 \times 0.75}{1000}}}$$

$$= \frac{0.05}{0.01837} = 2.72$$

Since $|Z| = 2.72 > 1.96$, null hypothesis is rejected at 5% level of significance. Hence at 5% level of significance these samples will exhibit the difference in the population proportions.

EXERCISE 5.3

1. A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.
2. In a hospital 525 female and 475 male babies were born in a month. Do these figures confirm the hypothesis that females and males are born in equal number ?
3. A die is thrown 10000 times and a throw of 3 or 4 was obtained 4200 times. On the assumption of random throwing do the data indicate an unbiased die ?
4. Given that on the average 4% of insured men of age 65 die within a year and that 60 of a particular group of 1000 such men (age 65) died within a year. Can this group be regarded as a representative sample ?
5. 325 men out of 600 men chosen from a big city were found to be smokers. Does this information support the conclusion that the majority of men in the city are smokers ?
6. A random sample of 400 apples is taken from a large basket and 40 are found to be bad. Estimate the proportion of bad apples in the basket and assign limits within which the percentage most probably lies.

NOTES

7. A manufacturer claimed that at least 95% of the equipments which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipments revealed that 18 were faulty. Test the manufacturer's claim at a level of significance (*i*) 5% (*ii*) 1%.
8. 1000 articles from a factory are examined and found to be 2.5% defective. 1500 similar articles from a second factory are found to have only 2% defectives. Can it reasonably concluded that the products of the first factory are inferior to those of second ?
9. A manufacturing firm claims that its brand A product outsells its brand B product by 8%. If it is found that 42 out of a sample of 200 persons prefer brand A and 18 out of another sample of 100 persons prefer brand B. Test whether the 8% difference is valid claim.
10. In a survey on a particular matter in a college, 850 males and 560 females voted. 500 males and 320 females voted yes. Does this indicate a significant difference of opinion between male and female on this matter at 1% level of significance ?
11. Two samples of sizes 1200 and 900 respectively drawn from two large populations. In the two large populations there are 30% and 25% respectively of fair haired people. Test whether these two samples will reveal the difference in the population proportions.
12. Before an increase in excise duty on tea 800 persons out of a sample of 1000 persons were found to be tea drinkers. After an increase in excise duty 800 people were tea drinkers in a sample of 1200 people. Test whether there is a significant decrease in the consumption of tea after the increase in excise duty.

Answers

1. H_0 is accepted at 5% level of significance
2. Yes, H_0 is accepted at 5% level of significance
3. H_0 is rejected
4. H_0 is rejected
5. H_0 is rejected at 5% level of significance
6. 8.5 : 11.5
7. Using left tailed test, H_0 is rejected at both 5% and 1% level of significance
8. No, H_0 is accepted
9. H_0 is accepted
10. H_0 is accepted
11. H_0 is rejected at 5% level of significance
12. H_0 is rejected.

5.17. TEST OF SIGNIFICANCE FOR SINGLE MEAN

This test is used to test the significant difference between sample mean and population mean.

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a normal population with mean μ and variance σ^2.

The standard error (S.E.) of mean of a random sample of size n from a population is given by

S.E. $(\bar{x}) = \dfrac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population.

We set the null hypothesis H_0 that the sample has been drawn from a large population with mean μ and variance σ^2 *i.e.*, there is no significant difference between the sample mean $(\bar{x})$ and population mean (μ).

Under the null hypothesis H_0 the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

NOTES

If standard deviation of the population (σ) is not known, we use the test statistic given as

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \text{ where } s \text{ is the standard deviation of the sample.}$$

Note. The limits of the population mean μ are given by $\bar{x} \pm Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$ *i.e.*,

$$\bar{x} - Z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

These limits are called the confidence limits for μ.

SOLVED EXAMPLES

Example 1. *A normal population has a mean of 6.8 and standard deviation of 1.5. A sample of 400 members gave a mean of 6.75. Is the difference significant ?*

Solution. Here, $\mu = 6.8$, $\bar{x} = 6.75$, $\sigma = 1.5$, $n = 400$

Null hypothesis H_0 : $\bar{x} = \mu$ (there is no significant difference between $\bar{x}$ and μ)

Alternative hypothesis H_1 : there is a significant difference between $\bar{x}$ and μ

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.75 - 6.8}{1.5/\sqrt{400}} = -\frac{0.05}{0.075} = -0.67$$

Since $|Z| = 0.67 < 1.96$ H_0 is accepted at 5% level of significance. Hence there is no significant difference between $\bar{x}$ and μ .

Example 2. *A random sample of 400 members has a mean 99. Can it be reasonably regarded as a sample from a large population of mean 100 and standard deviation 8 at 5% level of significance ?*

Solution. Here, $\mu = 100$, $\bar{x} = 99$, $\sigma = 8$, $n = 400$

Null hypothesis H_0 : the sample is drawn from a large population with mean 100 and standard deviation 8.

Alternative hypothesis H_1 : $\mu \neq 100$ (two tailed test)

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{99 - 100}{8/\sqrt{400}} = -\frac{1}{0.4} = -2.5$$

Since $|Z| = 2.5 > 1.96$, H_0 is rejected at 5% level of significance. Hence there is a significant difference between $\bar{x}$ and μ *i.e.*, it can not be regarded as a sample from a large population.

Example 3. *The management of a company claims that the average weekly income of their employees is ₹ 900. The trade union disputes this claim stressing that it is rather less. An independent sample of 150 randomly selected employees estimated the average to be ₹ 856 with standard deviation of ₹ 354. Would you accept the view of the management ?*

Solution. Here, $\mu = 900$, $\bar{x} = 854$, $s = 354$, $n = 150$

Null hypothesis H_0 : there is no significant difference between $\bar{x}$ and μ *i.e.*, the view of management is correct.

Alternative hypothesis H_1 : $\mu \neq 900$ (two-tailed test)

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{854 - 900}{354/\sqrt{150}} = -\frac{46}{28.904} = -1.59$$

Since $|Z| = 1.59 < 1.96$, H_0 is accepted at 5% level of significance. Hence the view of management is correct.

NOTES

Example 4. *In a population with a standard deviation of 14.8, what sample size is needed to estimate the mean of population within ± 1.2 with 95% confidence ?*

Solution. Here, $\bar{x} - \mu = \pm 1.2$, $\sigma = 14.8$, $Z = 1.96$

We know that $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Using this, we have

$$1.96 = \frac{\pm 1.2}{14.8/\sqrt{n}} = \frac{\pm 1.2\sqrt{n}}{14.8}$$

On squaring both the sides we have

$$(1.96)^2 = \left(\frac{\pm 1.2}{14.8}\right)^2 \times n \quad \text{or} \quad n = \left(\frac{1.96 \times 14.8}{\pm 1.2}\right)^2 = 584.35 \approx 584.$$

Example 5. *A random sample of 900 measurements from a large population gave a mean value of 64. If this sample has been drawn from a normal population with standard deviation of 20, find the 95% and 99% confidence limits for the mean in the population.*

Solution. Here, $n = 900$, $\bar{x} = 64$, $\sigma = 20$

At 95% confidence $Z = 1.96$

At 99% confidence $Z = 2.58$

The confidence limits for the population mean μ is given by

$$\bar{x} \pm Z_\alpha \frac{\sigma}{\sqrt{n}}$$

The confidence limits for 95% confidence are

$$64 \pm 1.96 \times \frac{20}{\sqrt{900}} = 64 \pm 1.307 = 62.693 \text{ and } 65.307$$

The confidence limits for 99% confidence are

$$64 \pm 2.58 \times \frac{20}{\sqrt{900}} = 64 \pm 1.72 = 62.28 \text{ and } 65.72.$$

EXERCISE 5.4

1. A random sample of 900 members has a mean 3.4 cms. Can it be reasonably regarded as a sample from a large population of mean 3.2 cms and standard deviation 2.3 cms ?
2. A random sample of 400 male students is found to have a mean height of 160 cms. Can it be reasonably regarded as a sample from a large population with mean height 162.5 cms and standard deviation 4.5 cms ?
3. A random sample of 200 measurements from a large population gave a mean value of 50 and a standard deviation of 9. Determine 95% confidence interval for the mean of population.
4. A random sample of 400 measurements from a large population gave a mean value of 82 and a standard deviation of 18. Determine 95% confidence interval for the mean of population.
5. A company manufacturing electric bulbs claims that the average life of its bulbs is 1600 hours. The average life and standard deviation of random sample of 100 such bulbs were 1570 hours and 120 hours respectively. Should we accept the claim of the company ?

NOTES

6. An insurance agent has claimed that the average age of policy holders who insure through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy holders who had insured through him reveal that the mean and standard deviation are 28.8 years and 6.35 years respectively. Test his claim at 5% level of significance.
7. The guaranteed average life of a certain type of bulbs is 1000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90% of the bulbs do not fall short of the guaranteed average by more than 2.5%. What must be the minimum size of the sample ?

Answers

1. Yes, H_0 is accepted
2. Yes, H_0 is accepted
3. 48.8 and 51.2
4. 80.24 and 83.76
5. No, rejected at 5% level of significance
6. Claim is valid
7. $n = 4$

5.18. TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

(*i*) This test is used to test the significant difference between the means of two large samples.

Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2.

We set the null hypothesis H_0 that there is no significant difference between the sample means *i.e.*, $\mu_1 = \mu_2$.

Under the null hypothesis H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If the samples are drawn from the same population with common standard deviation (σ), then under the null hypothesis the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (\because \quad \sigma_1 = \sigma_2 = \sigma)$$

Note. 1. If $\sigma_1 \neq \sigma_2$ and σ_1 and σ_2 are not known, the test statistic is

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

2. If common standard deviation (σ) is not known and $\sigma_1 = \sigma_2$ than σ can be obtained by using

$$\sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}$$

The test statistic is $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

NOTES

(*ii*) **Standard Deviations.** This test is used to test the significant difference between the standard deviations of two populations.

Let two independent random sample of sizes n_1 and n_2 having standard deviations s_1 and s_2 be drawn from the two normal population with standard deviation σ_1 and σ_2 respectively.

We set the null hypothesis H_0 that the sample standard deviations do not differ significantly *i.e.*, $\sigma_1 = \sigma_2$.

Under the null hypothesis H_0 the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

If σ_1 and σ_2 are unknown then the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}.$$

SOLVED EXAMPLES

Example 1. *Examine whether there is any significant difference between the two samples for the following data :*

Sample	*Size*	*Mean*
1	*50*	*140*
2	*60*	*150*

Standard deviation of the population = 10.

Solution. Here, $n_1 = 50$, $n_2 = 60$, $\bar{x}_1 = 140$, $\bar{x}_2 = 150$, $\sigma = 10$

Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, samples are drawn from the same normal population.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Under H_0 the test statistics is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{140 - 150}{10\sqrt{\frac{1}{50} + \frac{1}{60}}} = -\frac{10}{1.915} = -5.22$$

Since $|Z| = 5.22 > 3$, H_0 is rejected. Hence the samples are not drown from the same normal population.

NOTES

Example 2. *Intelligence tests on two groups of boys and girls gave the following results.*

	Mean	*S.D.*	*Size*
Girls	*70*	*10*	*70*
Boys	*75*	*11*	*100*

Examine if the difference between mean scores is significant.

Solution. Here, $n_1 = 70$, $n_2 = 100$, $\bar{x}_1 = 70$, $\bar{x}_2 = 75$, $s_1 = 10$, $s_2 = 11$

Null hypothesis H_0 : There is no significant difference between mean scores *i.e.,* $\bar{x}_1 = \bar{x}_2$

Alternative hypothesis H_1 : $\bar{x}_1 \neq \bar{x}_2$ (two-tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{70 - 75}{\sqrt{\frac{10^2}{70} + \frac{11^2}{100}}} = -\frac{5}{2.639} = -1.895$$

Since $|Z| = 1.895 < 1.96$, H_0 is accepted at 5% level of significance. Hence there is no significant difference between mean scores.

Example 3. *Two samples were taken from two normal populations. The following information was available on these samples regarding the expenditure in Rupees per month per family.*

Sample 1 $n_1 = 42$ $\bar{x}_1 = 744.85$ $\sigma_1^2 = 158165.43$

Sample 2 $n_2 = 32$ $\bar{x}_2 = 516.78$ $\sigma_2^2 = 26413.61$

Test whether the average expenditure per month per family is equal.

Solution. Null hypothesis H_0 : $\mu_1 = \mu_2$ *i.e.,* the average expenditure per month per family is equal.

Alternative hypothesis H_1 : $\mu_1 \neq \mu_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{744.85 - 516.78}{\sqrt{\frac{158165.43}{42} + \frac{26413.61}{32}}} = \frac{228.07}{67.76} = 3.37$$

Since $|Z| = 3.37 > 1.96$, H_0 is rejected at 5% level of significance. Hence the average expenditure per month per family is not equal.

Example 4. *The means of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can the samples be regarded as drawn from the same population of standard deviation 6.25 cms ?*

Solution. Here, $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 168.75$, $\bar{x}_2 = 170$, $\sigma = 6.25$

Null hypothesis H_0 : $\mu_1 = \mu_2$ *i.e.*, samples are drawn from the same population.

Alternative hypothesis H_1 : $\mu_1 \neq \mu_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{168.75 - 170}{6.25\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -\frac{1.25}{0.242} = -5.165$$

NOTES

Since $|Z| = 5.165 > 1.96$, H_0 is rejected at 5% level of significance. Hence the samples are not drawn from the same population.

Example 5. *Two random samples of sizes 1000 and 2000 farms gave an average yield of 2000 kg and 2050 kg respectively. The variance of wheat farms in the country may be taken as 10 kg. Examine whether the two samples differ significantly in yield.*

Solution. Here, $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 2000$, $\bar{x}_2 = 2050$, $\sigma^2 = 100$ *i.e.*, $\sigma = 10$

Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, samples are drawn from the same population.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2000 - 2050}{10\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -\frac{50}{0.387} = -129.20$$

Since $|Z| = 129.20 > 3$ (maximum value of Z), highly significant, H_0 is rejected. Hence the samples are not drawn from the same normal population.

Example 6. *The standard deviation of weight of all students in a college was found to be 4 kgs. Two random samples are drawn. The standard deviations of the weight of 100 undergraduate students is 3.5 kgs and 50 postgraduate students is 3 kgs. Test the significance of the difference in standard deviations of the samples at 5% level of significance.*

Solution. Here, $n_1 = 100$, $n_2 = 50$, $s_1 = 3.5$, $s_2 = 3$, $\sigma = 4$

Null hypothesis $H_0 : \sigma_1 = \sigma_2$ *i.e.*, sample standard deviations do not differ significantly

Alternative hypothesis $H_1 : \sigma_1 \neq \sigma_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{s_1 - s_2}{\sigma\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}} = \frac{3.5 - 3}{4\sqrt{\frac{1}{200} + \frac{1}{100}}} = \frac{0.5}{0.49} = 1.02$$

Since $|Z| = 1.02 < 1.96$, H_0 is accepted. Hence sample standard deviations do not differ significantly.

Example 7. *Random samples drawn from two large cities gave the following information relating to the heights of adult males :*

	Mean height (in inches)	*Standard deviation*	*No. in samples*
City 1	*67.42*	*2.58*	*1000*
City 2	*67.25*	*2.50*	*1200*

Test the significance of difference in standard deviations of the samples at 5% level of significance.

NOTES

Solution. Here, $n_1 = 1000$, $n_2 = 1200$, $\bar{x}_1 = 67.42$, $\bar{x}_2 = 67.25$, $s_1 = 2.58$, $s_2 = 2.50$, σ is not known.

Null hypothesis $H_0 : \sigma_1 = \sigma_2$ *i.e.*, the sample standard deviations do not differ significantly

Alternative hypothesis $H_1 : \sigma_1 \neq \sigma_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2000} + \frac{(2.50)^2}{2400}}} = \frac{0.08}{0.077} = 1.039$$

Since $|Z| = 1.039 < 1.96$, H_0 is accepted. Hence sample standard deviations do not differ significantly.

Example 8. *In a survey of incomes of two classes of workers of two random samples gave the following data :*

	Size of sample	*Mean annual income in* ₹	*Standard deviation in* ₹
Sample 1	*100*	*582*	*24*
Sample 2	*100*	*546*	*28*

Examine whether the difference between

(*i*) *Mean and*

(*ii*) *The standard deviations significant.*

Solution. Here, $n_1 = 100$, $n_2 = 100$, $\bar{x}_1 = 582$, $\bar{x}_2 = 546$, $s_1 = 24$, $s_2 = 28$

(*i*) Null hypothesis $H_0 : \mu_1 = \mu_2$ *i.e.*, sample means do not differ significantly.

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (two tailed test)

Under H_0 the test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{582 - 546}{\sqrt{\frac{(24)^2}{100} + \frac{(28)^2}{100}}} = \frac{36}{3.6878} = 9.762$$

Since $|Z| = 9.762 > 1.96$, highly significant, H_0 is rejected at 5% level of significance. Hence sample means differ significantly.

(*ii*) Null hypothesis $H_0 : \sigma_1 = \sigma_2$ *i.e.,* sample standard deviations do not differ significantly.

Alternative hypothesis $H_1 : \sigma_1 \neq \sigma_2$ (two-tailed test)

Under H_0 the test statistic is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{24 - 28}{\sqrt{\frac{(24)^2}{200} + \frac{(28)^2}{200}}} = \frac{-4}{2.6077} = -1.53$$

Since $|Z| = 1.53 < 1.96$, H_0 is accepted at 5% level of significance. Hence sample standard deviations do not differ significantly.

NOTES

EXERCISE 5.5

1. The number of accidents per day were studied for 144 days in city A and for 100 days in city B. The mean numbers of accidents and standard deviations were respectively 4.5 and 1.2 for city A and 5.4 and 1.5 for city B. Is city A more prone to accidents than city B.

2. The mean yields of a crop from two places in a district were 210 kgs and 220 kgs per acre from 100 acres and 150 acres respectively. Can it be regarded that the sample were drawn from the same district which has the standard deviation of 11 kgs per acre ?

3. Given the following data :

	No. of cases	*Mean wages in* ₹	*Standard deviation of wages in* ₹
Sample 1	400	47.4	3.1
Sample 2	900	50.3	3.3

Examine whether the two mean wages differ significantly.

4. A sample of heights of 6400 soldiers has a mean of 67.85 inches and a standard deviation of 2.56 inches. While another sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers ?

5. Intelligence tests on two groups of boys and girls gave the following results :

	Mean	*S.D*	*Size*
Girls	75	8	60
Boys	73	10	100

Examine if the difference between mean scores is significant.

6. The yield of a crop in a random sample of 1000 farms in a certain area has a standard deviation of 192 kgs. Another random sample of 1000 farms gives a standard deviation of 224 kgs. Are the standard deviations significantly different ?

7. The standard deviation of a random sample of 900 members is 4.6 and that of another random sample of 1600 is 4.8. Examine if the standard deviations are significantly different.

8. The mean yield of two sets of plots and their variability are as follows :

	Set of 40 plots	*Set of 60 plots*
Mean yield per plot	1258 kgs	1243 kgs
S.D. per plot	34	28

Examine whether

(*i*) the difference in the variability in yields is significant,

(*ii*) the difference in the mean yields is significant.

Answers

1. No **2.** No **3.** Yes, highly significant

4. Highly significant **5.** Not significant at 5% **6.** Yes

7. Not significant **8.** (*i*) Not significant (*ii*) significant.

NOTES

6. NON-PARAMETRIC TESTS

STRUCTURE

Chi-square Test
Chi-square Test to Test the Goodness of Fit
Chi-square Test to Test the Independence of Attributes
Conditions for χ^2 Test
Uses of χ^2 Test
Correlation Analysis
Scatter or Dot Diagram
Characteristics of the coefficient of Correlation *r*
Spearman's Rank Correlation

6.1. CHI-SQUARE TEST

In test of hypothesis of parameters, it is usually assumed that the random variable follows a particular distribution. To confirm whether our assumption is right, Chi-square test is used which measures the discrepancy between the observed (actual) frequencies and theoretical (expected) frequencies, on the basis of outcomes of a trial or observational data. Chi-square is a letter of the Greek alphabet and is denoted by χ^2. It is a continuous distribution which assumes only positive values.

6.2. CHI-SQUARE TEST TO TEST THE GOODNESS OF FIT

The value of χ^2 is used to test whether the deviations of the observed (actual) frequencies from the theoretical (expected) frequencies are significant or not. Chi-square test is also used to test whether a set of observations fit a given distribution or not. Therefore, chi-square provides a test of goodness of fit.

If $O_1, O_2, \ldots\ldots, O_n$ is a set of observed (actual) frequencies and $E_1, E_2, \ldots\ldots, E_n$ is the corresponding set of theoretical (expected) frequencies, then the statistic χ^2 is given by

$$\chi^2 = \sum_{i=1}^{n} \left\{ \frac{(O_i - E_i)^2}{E_i} \right\}$$

is distributed with $(n - 1)$ degrees of freedom.

NOTES

Here, we test the null hypothesis

H_0 : There is no significant difference between the observed (actual) values and the corresponding expected (theoretical) values.

v.s., H_1 : H_0 is not true.

If $\chi^2_{cal} \geq \chi^2_{tab}$ (or $\chi^2_{\alpha,\, n-1}$) then H_0 is rejected otherwise H_0 is accepted.

Note. If the null hypothesis H_0 is true, the test statistic χ^2 follow chi-square distribution with $(n - 1)$ degrees of freedom, where

$$\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i \; ; \quad i.e. \quad \sum_{i=1}^{n} (O_i - E_i) = 0.$$

6.3. CHI-SQUARE TEST TO TEST THE INDEPENDENCE OF ATTRIBUTES

The value of χ^2 is used to test whether two attributes are associated or not, *i.e.* independence of attributes. To test the independence of attributes contingency table is used.

A contingency table is a two-way table in which rows are classified according to one attribute or criterion and columns are classified according to the other attribute or criterion. Each cell contains that number of items O_{ij} possessing the qualities of the ith row and jth column, where $i = 1, 2, \ldots\ldots, r$ and $j = 1, 2, \ldots\ldots, s$. In such a case contingency table is said to be of order $(r \times s)$. Each row or column total is known as marginal total. Also we have the sum of row totals $\sum_{i=1}^{r} R_i$ is equal to the sum of column totals $\sum_{j=1}^{s} C_j$, *i.e.*

$$\sum_{i} R_i = \sum_{j} C_j = N, \text{ where } N \text{ is the total frequency.}$$

Let us consider the two attributes A and B, where A divided into r classes $A_1, A_2, \ldots\ldots, A_r$ and B divided into s classes $B_1, B_2, \ldots\ldots, B_s$. If R_i represents the number of persons possessing the attributes A_i ; C_j represents the number of persons possessing

NOTES

the attributes B_j and O_{ij} represent the number of persons possessing attributes A_i and B_j respectively. The contingency table of order $(r \times s)$ is shown in the following table :

Columns / Rows	B_1	B_2	.....	B_s	Total
A_1	O_{11}	O_{12}	.....	O_{1s}	R_1
A_2	O_{21}	O_{22}	.....	O_{2s}	R_2
⋮	⋮	⋮		⋮	⋮
A_r	O_{r1}	O_{r2}	.....	O_{rs}	R_r
Total	C_1	C_2	.....	C_s	N

Corresponding to each O_{ij} the expected frequency E_{ij} in a contingency table is calculated by

$$E_{ij} = \frac{R_i \times C_j}{N} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Here, we test the null hypothesis

H_0 : There is no association between the attributes under study, *i.e.* attributes A and B are independent.

v.s., H_1 : attributes are associated, *i.e.,* attributes A and B are not independent.

H_0 can be tested by the statistic

$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is distributed with $(r-1)(s-1)$ degrees of freedom.

If ${\chi_{cal}}^2 \geq {\chi^2}_{tab}$ (or $\chi^2_{\alpha,(r-1)(s-1)}$), then H_0 is rejected otherwise H_0 is accepted.

Note 1. For a contingency table with r rows and s columns, the degrees of freedom $= (r-1)(s-1)$.

2. For a 2×2 contingency table $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ we use the following formula to calculate the value of statistic χ^2 as

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(b+d)(a+c)(c+d)},$$

where $N = a + b + c + d$

χ^2 has $(2-1)(2-1) = 1$ degree of freedom.

3. Yate's correction. In a 2×2 contingency table, if any of cell frequency is less than 5, we make a correction to make χ^2 continuous. Decrease by $\frac{1}{2}$ those cell frequencies which are greater than expected frequencies and increase by $\frac{1}{2}$ those cell frequencies which are less than expected frequencies. This will affect the marginal totals. This correction is known a Yate's correction.

After applying the Yate's correction, the corrected value of χ^2 is given by

$$\chi^2 = \frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(b+d)(a+c)(c+d)}.$$

6.4. CONDITIONS FOR χ^2 TEST

NOTES

1. The number of observations collected must be large, *i.e.* $n \geq 30$.
2. No theoretical frequency should be very small.
3. The sample observations should be independent.
4. N, the total of frequencies should be reasonably large, say, greater than 50.

6.5. USES OF χ^2 TEST

1. To test the goodness of fit.
2. To test the discrepancies between observed and expected frequencies.
3. To determine the association between attributes.

SOLVED EXAMPLES

Example 1. *The following table gives the number of accidents that took place in an industry during various days of the week. Test whether the accidents are uniformly distributed over the week.*

Days	*Mon.*	*Tue.*	*Wed.*	*Thu.*	*Fri.*	*Sat.*
No. of accidents	*16*	*20*	*14*	*13*	*17*	*16*

Solution. Here, $n = 6$, total number of accidents = 96

Null hypothesis H_0 : the accidents are uniformly distributed over the week.

Under H_0, the expected number of accidents of each of these days

$$= \frac{\text{Total no. of accidents}}{\text{No. of days}} = \frac{96}{6} = 16$$

The observed and expected number of accidents are given below :

O_i	16	20	14	13	17	16
E_i	16	16	16	16	16	16
$(O_i - E_i)^2$	0	16	4	9	1	0

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{0 + 16 + 4 + 9 + 1 + 0}{16} = \frac{30}{16} = 1.875.$$

Tabulated value of χ^2 for 5 (6 – 1 = 5) degrees of freedom at 5% level of significance is 11.07.

Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.,* the accidents are uniformly distributed over the week.

NOTES

Example 2. *A die is thrown 120 times and the result of these throws are given as :*

No. appeared on the die	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*16*	*30*	*22*	*18*	*14*	*20*

Test whether the die is biased or not.

Solution. Here, $n = 6$, total frequency = 120

Null hypothesis H_0 : die is unbiased

Under H_0, the expected frequencies for each digit = $\frac{120}{6} = 20$

The observed and expected frequencies are given below :

O_i	16	30	22	18	14	20
E_i	20	20	20	20	20	20
$(O_i - E_i)^2$	16	100	4	4	36	0

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{16 + 100 + 4 + 4 + 36 + 0}{20} = \frac{160}{20} = 8$$

Tabulated value of χ^2 for 5 (6 – 1 = 5) degrees of freedom at 5% level of significance is 11.07. Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.* the die is unbiased.

Example 3. *The following table shows the distribution of digits in numbers chosen at random from a telephone directory :*

Digits	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Frequency	*1026*	*1107*	*997*	*966*	*1075*	*933*	*1107*	*972*	*964*	*853*

Test at 5% level whether the digits may be taken to occur equally frequently in the directory.

Solution. Here, $n = 10$, total frequency = 10,000

Null hypothesis H_0 : all the digits occur equally frequently in the directory

Under H_0, the expected frequency of each of the digits = $\frac{10{,}000}{10} = 1000$

The observed and expected frequencies are given below :

O_i	1026	1107	997	966	1075	933	1107	972	964	853
E_i	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
$(O_i - E_i)^2$	676	11449	9	1156	5625	4489	11449	784	1296	21609

$$\chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = \frac{676 + 11449 + \ldots\ldots + 21609}{1000}$$

$$= \frac{58542}{1000} = 58.542$$

Tabulated value of χ^2 for 9 (10 – 1 = 9) degrees of freedom at 5% level of significance is 16.92.

Since calculated value of χ^2 is greater than tabulated value of χ^2, so H_0 is rejected, *i.e.* all the digits in the numbers in the telephone directory do not occur equally frequently.

NOTES

Example 4. *Survey of 320 families of 5 children each revealed the following information :*

No. of male births	*5*	*4*	*3*	*2*	*1*	*0*
No. of female births	*0*	*1*	*2*	*3*	*4*	*5*
No. of families	*14*	*56*	*110*	*88*	*40*	*12*

Test whether the data are consistent with the hypothesis that Binomial law holds and the chance of male and female births are equally probable.

Solution. Null hypothesis

H_0 : The male and female births are equally probable *i.e.*, $p = q = \frac{1}{2}$, where p is the probability of female birth and q is the probability of male birth.

The expected frequencies are calculated by using Binomial distribution as :

$E(r) = N \times P\ (X = r)$, where $r = 0, 1, 2, 3, 4, 5$; where N is the total frequency and $E(r)$ is the number of families with r female children.

$$P(X = r) = {}^{n}C_r\, p^r q^{n-r}\ ;\ n \text{ is number of children.}$$

E(0) = No. of families with 0 female children

$$= 320 \times {}^{5}C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{5-0} = 320 \times \frac{1}{32} = 10$$

E(1) = No. of families with 1 female children

$$= 320 \times {}^{5}C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{5-1} = 320 \times \frac{5}{32} = 50$$

E(2) = No. of families with 2 female children

$$= 320 \times {}^{5}C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{5-2} = 320 \times \frac{10}{32} = 100$$

E(3) = No. of families with 3 female children

$$= 320 \times {}^{5}C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{5-3} = 320 \times \frac{10}{32} = 100$$

E(4) = No. of families with 4 female children

$$= 320 \times {}^{5}C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{5-4} = 320 \times \frac{5}{32} = 50$$

E(5) = No. of families with 5 female children

$$= 320 \times {}^{5}C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{5-5} = 320 \times \frac{1}{32} = 10$$

O_i	14	56	110	88	40	12
E_i	10	50	100	100	50	10
$(O_i - E_i)^2$	16	36	100	144	100	4

$$\chi^2 = \sum_{i=0}^{5} \frac{(O_i - E_i)^2}{E_i} = \frac{16}{10} + \frac{36}{50} + \frac{100}{100} + \frac{144}{100} + \frac{100}{50} + \frac{4}{10}$$

$$= 1.60 + 0.72 + 1.00 + 1.44 + 2.00 + 0.40 = 7.16$$

NOTES

Tabulated value of χ^2 for 5 (6 – 1 = 5) degrees of freedom at 5% level of significance is 11.07. Since calculated value of χ^2, is less than tabulated value of χ^2, so H_0 is accepted *i.e.*, male and female births are equally probable.

Example 5. *Fit a Poisson distribution for the following data and test the goodness of fit.*

No. of defects (x)	*0*	*1*	*2*	*3*	*4*	*5*
Frequency	*6*	*13*	*13*	*8*	*4*	*3*

Solution. Null hypothesis H_0 : Poisson distribution is a good fit to the data.

We first find the Poisson distribution for the above data.

Mean of given distribution = $\frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{94}{47} = 2$

Here, $\lambda = 2$ (For a Poisson distribution mean = λ)

$N = \Sigma f_i = 47$

The expected frequencies of the Poisson distribution are given by

$$E(r) = N \times e^{-\lambda} \frac{\lambda^r}{r!} = 47 \times e^{-2} \frac{2^r}{r!} \; ; r = 0, 1, 2, 3, 4, 5$$

The expected frequencies are as :

$$E(0) = 47 \times e^{-2} \cdot \frac{2^0}{0!} = 6.36 \approx 6 \qquad (e^{-2} = 0.1353)$$

$$E(1) = 47 \times e^{-2} \cdot \frac{2^1}{1!} = 12.72 \approx 13$$

$$E(2) = 47 \times e^{-2} \cdot \frac{2^2}{2!} = 12.72 \approx 13$$

$$E(3) = 47 \times e^{-2} \cdot \frac{2^3}{3!} = 8.48 \approx 9$$

$$E(4) = 47 \times e^{-2} \cdot \frac{2^4}{4!} = 4.24 \approx 4$$

$$E(5) = 47 \times e^{-2} \cdot \frac{2^5}{5!} = 1.696 \approx 2$$

x	0	1	2	3	4	5
O_i	6	13	13	8	4	3
E_i	6.36	12.72	12.72	8.48	4.24	1.696
$(O_i - E_i)^2$	0.1296	0.0784	0.0784	0.2304	0.0576	1.7004

$$\chi^2 = \sum_{i=0}^{5} \frac{(O_i - E_i)^2}{E_i} = \frac{0.1296}{6.36} + \frac{0.0784}{12.72} + \frac{0.0784}{12.72} + \frac{0.2304}{8.48} + \frac{0.0576}{4.24} + \frac{1.7004}{1.696}$$

$$= 0.02038 + 0.00616 + 0.00616 + 0.02717 + 0.01358 + 1.0026$$

$$= 1.07605$$

Tabulated value of χ^2 for 4 (6 – 2 = 4) degrees of freedom at 5% level of significance is 9.488.

Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.*, Poisson distribution is a good fit to the data.

NOTES

Example 6. *The theory predicts that the proportion of beans in the four groups A, B, C and D should be in the ratio 11 : 4 : 3 : 2. In an experiment with 2000 beans the number of four groups A, B, C and D are 1070, 430, 330 and 170 respectively. Does the experimental result support the theory.*

Solution. Null hypothesis H_0 : the experimental result support the theory, *i.e.* there is no significant difference between observed and theoretical frequencies.

Under H_0 the expected (theoretical) frequencies can be calculated as :

Total number of beans = 1070 + 430 + 330 + 170 = 2000

Sum of ratios = 11 + 4 + 3 + 2 = 20

$$E(A) = 2000 \times \frac{11}{20} = 1100$$

$$E(B) = 2000 \times \frac{4}{20} = 400$$

$$E(C) = 2000 \times \frac{3}{20} = 300$$

$$E(D) = 2000 \times \frac{2}{20} = 200$$

O_i	1070	430	330	170
E_i	1100	400	300	200
$(O_i - E_i)^2$	900	900	900	900

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{900}{1100} + \frac{900}{400} + \frac{900}{300} + \frac{900}{200}$$

$$= 0.8182 + 2.250 + 3.000 + 4.500 = 10.5682$$

Tabulated value of χ^2 for 3 (4 – 1 = 3) degrees of freedom at 5% level of significance is 7.815. Since calculated value of χ^2 is greater than tabulated value of χ^2, so H_0 is rejected, *i.e.* the experimental results does not support the theory.

Example 7. *Find the expected frequencies of 2 × 2 contingency table* $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$.

Solution.

Attributes	B_1	B_2	*Total*
A_1	a	b	$a + b$
A_2	c	d	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$

The expected frequencies are

$$E(a) = E(A_1, B_1) = \frac{(a+b)(a+c)}{a+b+c+d}$$

$$E(b) = E(A_1, B_2) = \frac{(a+b)(b+d)}{a+b+c+d}$$

NOTES

$$E(c) = E(A_2, B_1) = \frac{(c+d)(a+c)}{a+b+c+d}$$

$$E(d) = E(A_2, B_2) = \frac{(c+d)(b+d)}{a+b+c+d}$$

Example 8. *The following data is collected on two characters* :

	Smokers	*Non smokers*
Literate	*83*	*57*
Illiterate	*45*	*68*

From this information find out whether there is any relation between literacy and the smoking.

Solution. Null hypothesis H_0 : There is no relation between literacy and the smoking, *i.e.* they are independent

	Smokers	*Non smokers*	*Total*
Literate	83	57	140 (R_1)
Illiterate	45	68	113 (R_2)
Total	128 (C_1)	125 (C_2)	N = 253

Under the null hypothesis, expected frequencies can be calculated by using

$$E_{ij} = \frac{R_i \times C_j}{N} \quad (i = 1, 2 \; ; j = 1, 2)$$

Expected frequencies are

	Smokers	*Non smokers*	*Total*
Literate	$\frac{140 \times 128}{253} = 70.83$	$\frac{140 \times 125}{253} = 69.17$	140
Illiterate	$\frac{113 \times 128}{253} = 57.17$	$\frac{113 \times 125}{253} = 55.83$	113
Total	128	125	253

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(83 - 70.83)^2}{70.83} + \frac{(57 - 69.17)^2}{69.17} + \frac{(45 - 57.17)^2}{57.17} + \frac{(68 - 55.83)^2}{55.83}$$

$$= 2.091 + 2.141 + 2.590 + 2.653 = 9.475$$

Tabulated value of χ^2 for 1 [(2 – 1) (2 – 1) = 1] degree of freedom at 5% level of significance is 3.841.

Since calculated value of χ^2 is greater than tabulated value of χ^2, so H_0 is rejected, *i.e.*, there is a relation between literacy and smoking or they are not independent.

NOTES

Example 9. *In a locality 100 persons were randomly selected and asked about their educational achievements. The results are given below :*

Sex	*Education*		
	Middle	*High school*	*College*
Male	*10*	*15*	*25*
Female	*25*	*10*	*15*

Based on this information can you say the education depends on sex.

Solution. Null hypothesis H_0 : Education is independent of sex.

Under the null hypothesis expected frequencies can be calculated by using

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

$(i = 1, 2 ; j = 1, 2, 3)$

Sex	*Education*			
	Middle	*High school*	*College*	*Total*
Male	10	15	25	50 (R_1)
Female	25	10	15	50 (R_2)
Total	35(C_1)	25 (C_2)	40 (C_3)	N = 100

Expected frequencies are

Sex	*Education*			
	Middle	*High school*	*College*	*Total*
Male	$\frac{50 \times 35}{100} = 17.5$	$\frac{50 \times 25}{100} = 12.5$	$\frac{50 \times 40}{100} = 20$	50
Female	$\frac{50 \times 35}{100} = 17.5$	$\frac{50 \times 25}{100} = 12.5$	$\frac{50 \times 40}{100} = 20$	50
Total	35	25	40	100

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(10 - 17.5)^2}{17.5} + \frac{(15 - 12.5)^2}{12.5} + \frac{(25 - 20)^2}{20} + \frac{(25 - 17.5)^2}{17.5} + \frac{(10 - 12.5)^2}{12.5} + \frac{(15 - 20)^2}{20}$$

$$= 3.214 + 0.5 + 1.25 + 3.214 + 0.5 + 1.25 = 9.928$$

Tabulated value of χ^2 for 2 [(2 – 1) (3 – 1) =2] degrees of freedom at 5% level of significance is 5.991. Since calculated value of χ^2 is greater than tabulated value of χ^2, so H_0 is rejected, *i.e.*, education is not independent of sex or there is a relation between education and sex.

NOTES

Example 10. *From the following table regarding the colour of eyes of father and son, test whether the colour of the son's eyes is associated with that of father's.*

Eye colour of father	*Eye colour of son*		*Total*
	Light	*Not light*	
Light	*471*	*151*	*622*
Not light	*148*	*230*	*378*
Total	*619*	*381*	*1000*

Solution. Null hypothesis H_0 : The colour of son's eye is not associated with that of father, *i.e.,* they are independent.

Under the null hypothesis expected frequencies can be calculated by using

$$E_{ij} = \frac{R_i \times C_j}{N} \quad (i = 1, 2 ;\ j = 1, 2)$$

Expected frequencies are

Eye colour of father	*Eye colour of son*		*Total*
	Light	*Not light*	
Light	$\frac{622 \times 619}{1000} = 385.018$	$\frac{622 \times 381}{1000} = 236.982$	622
Not light	$\frac{378 \times 619}{1000} = 233.982$	$\frac{378 \times 381}{1000} = 144.018$	378
Total	619	381	1000

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(471 - 385.018)^2}{385.018} + \frac{(151 - 236.982)^2}{236.982} + \frac{(148 - 233.982)^2}{233.982} + \frac{(230 - 144.018)^2}{144.018}$$

$$= 19.201 + 31.196 + 31.596 + 51.333 = 133.326$$

Tabulated value of χ^2 for 1 [(2 – 1) (2 – 1) = 1] degree of freedom at 5% level of significance is 3.841.

Since calculated value of χ^2 is greater than tabulated value of χ^2, so H_0 is rejected, *i.e.,* the colour of son's eye is associated with that of father or they are dependent.

Example 11. *The following table gives the number of good and bad parts produced by each of the three shifts in a factory.*

	Good parts	*Bad parts*	*Total*
Day shift	*960*	*40*	*1000*
Evening shift	*940*	*50*	*990*
Night shift	*950*	*45*	*995*
Total	*2850*	*135*	*2985*

NOTES

Test whether the production of bad parts is independent of the shifts on which they were produced.

Solution. Null hypothesis H_0 : The production of bad parts is independent of the shift on which they were produced, *i.e.* production and shifts are independent.

Under the null hypothesis expected frequencies can be calculated by using

$$E_{ij} = \frac{R_i \times C_j}{N} \quad (i = 1, 2, 3 ; j = 1, 2)$$

Expected frequencies are

	Good parts	*Bad parts*	*Total*
Day shift	$\frac{1000 \times 2850}{2985} = 954.774$	$\frac{1000 \times 135}{2985} = 45.226$	1000
Evening shift	$\frac{990 \times 2850}{2985} = 945.226$	$\frac{990 \times 135}{2985} = 44.774$	990
Night shift	$\frac{995 \times 2850}{2985} = 950.000$	$\frac{995 \times 135}{2985} = 45.000$	995
	2850	135	2985

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(960 - 954.774)^2}{954.774} + \frac{(40 - 45.226)^2}{45.226} + \frac{(940 - 945.226)^2}{945.226}$$

$$+ \frac{(50 - 44.774)^2}{44.774} + \frac{(950 - 950)^2}{950} + \frac{(45 - 45)^2}{45}$$

$$= 0.0286 + 0.6039 + 0.0289 + 0.6099 + 0 + 0 = 1.2713$$

Tabulated value of χ^2 for 2 [(3 – 1) (2 – 1) = 2] degrees of freedom at 5% level of significance is 5.991

Since calculated value of χ^2 is less than tabulated value of χ^2, so H_0 is accepted, *i.e.*, the production of bad parts is independent of the shift on which they were produced.

EXERCISE 6.1

1. The frequency distribution of the digits on a set of random numbers was observed to be :

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Test the hypothesis that the digits are uniformly distributed.

2. The sales in a supermarket during a week are given below :

Days	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
(*Sales ,000* ₹)	65	54	60	56	71	84

Test the hypothesis that the sales do not depend on the day of the week, using a 5% significant level.

NOTES

3. The following table gives the number of accidents that took place in an industry during various days of the week :

Days	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
No. of accidents	14	18	12	11	15	14

Test if accidents are uniformly distributed over the week.

4. A die is thrown 276 times and the results of these throws are given below :

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	29	59	57	59

Test whether the die is biased or not.

5. A sample analysis of examination results of 500 students was made. It was found that 220 had failed ; 170 had secured a third class ; 90 were placed in second class ; 20 got first class. Are these results commensurable with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for the above said categories respectively.

6. Four dice were thrown 112 times and the number of times 1, 3 or 5 was thrown were as under :

No. of dice throwing 1, 3 or 5	0	1	2	3	4
Frequency	10	25	40	30	7

Test the hypothesis that all dice were fair.

7. Fit a Poisson distribution for the following data and test the goodness of fit.

No. of defects (x)	0	1	2	3	4
Frequency	109	65	22	3	1

8. The following table gives the classification of 100 workers according to sex and nature of work. Using χ^2-test examine whether the nature of work is independent of the sex of the worker.

Sex	*Nature of work*	
	Skilled	*Unskilled*
Male	40	20
Female	10	30

9. For the data given in the following table use χ^2 -test to test the effectiveness of inoculation in preventing the attack of smallpox.

	Attacked	*Not attacked*
Inoculated	25	220
Not inoculated	90	160

10. Two investigators draw samples from the same town in order to estimate the number of persons falling in the income groups 'poor', middle class' and 'well to do'. Their results are as follows :

Investigator	*Income groups*		
	Poor	*Middle class*	*Well to do*
A	140	100	15
B	140	50	20

Test whether the sampling techniques of the two investigators are significantly dependent of the income groups of people.

Answers

1. Yes
2. Accepted
3. Yes
4. Biased
5. No
6. Yes
7. Poisson distribution is a good fit to the data
8. No
9. Inoculation against smallpox is a preventive measure
10. Sampling techniques are dependent of the income groups

6.6. CORRELATION ANALYSIS

In a bivariate distribution, if the change in one variable is accompanied by a change in the other variable in such a way that an increase in one variable results in an increase or decrease in the other, then the two variables are said to be correlated. For example, income and expenditure, heights and weights of students in a class, price and demand of certain commodities.

If the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive. But if the increase (or decrease) in one variable results in a corresponding decrease (or increase), in the other, correlation is said to be negative. If two variables vary in such a way that their ratio is always constant, then the correlation is said to be perfect.

6.7. SCATTER OR DOT DIAGRAM

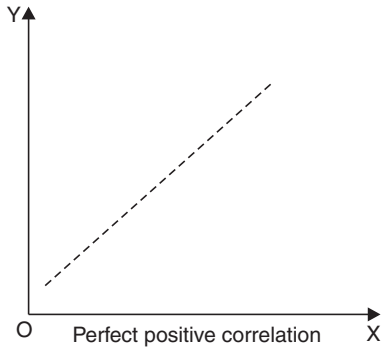
When we plot the corresponding values of two variables, taking one on X-axis and the other along Y-axis, it shows a collection of dots. This collection of dots is called a dot diagram or a scatter diagram.

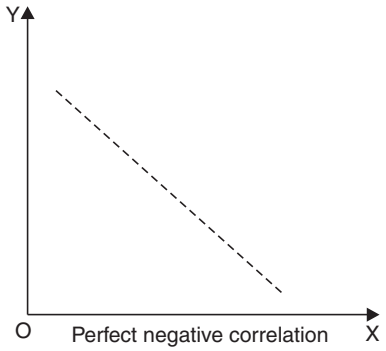
If all the plotted points lie in a straight line and show an upward trend, then the correlation is perfect positive. If all the plotted points lie in a straight line and show a downward trend, then the correlation is perfect negative.

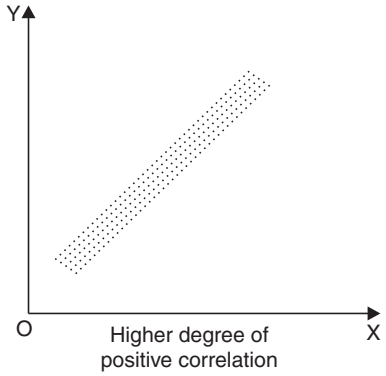
If the plotted points are not on a straight line but seem to be scattered around a straight line, the variables are correlated. Closer the scatter of points around a line, higher is the degree of correlation. If the plotted points are not clustered around a straight line but are widely scattered over the diagram, then there is a very low degree of correlation between the variables.

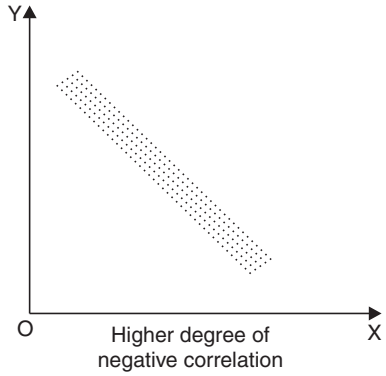
If the plotted points show no trend at all, then the variables are independent and are not correlated.

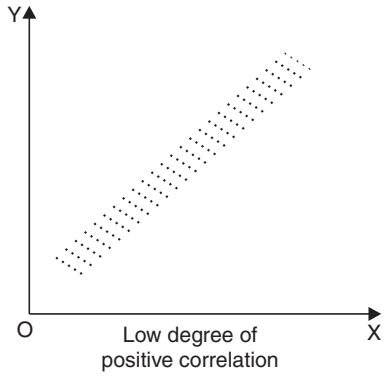
NOTES

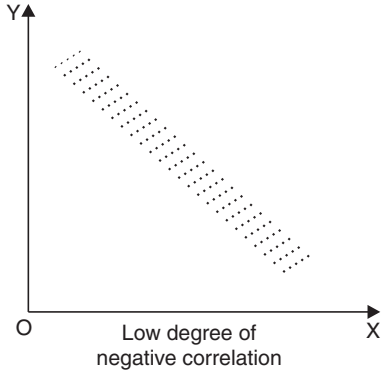


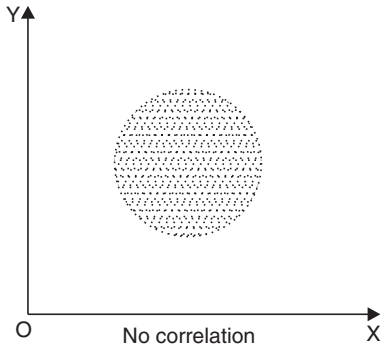












Karl Pearson's Coefficient of Correlation

NOTES

The correlation coefficient $r\ (x, y)$ between two variables x and y is given by

$$r\ (x, y) = \frac{\text{cov}\ (x, y)}{\sqrt{\text{variance}\ (x)}\ \sqrt{\text{variance}\ (y)}} = \frac{\text{cov}\ (x, y)}{\sigma_x\ \sigma_y}$$

$r\ (x, y)$ is also denoted by $\rho\ (x, y)$ or r_{xy} or simply by r.

$$r(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

or

$$r(x, y) = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

If the values of x_i or y_i's are large or involve fractions, then define

$$u_i = \frac{x_i - a}{h} \quad \text{and} \quad v_i = \frac{y_i - b}{k},$$

where a and b are assumed means of x-series and y-series respectively, h and k are constants. This property is known as change of origin and scale. Correlation coefficient is independent of change of origin and scale. In this $r(x, y)$ is given by the formula.

$$r(x, y) = r(u, v) = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2}}.$$

6.8. CHARACTERISTICS OF THE COEFFICIENT OF CORRELATION *r*

(*i*) $-1 \le r \le +1$

(*ii*) If $r = -1$, then there is perfect negative correlation between x and y.

(*iii*) If $r = 1$, then there is perfect positive correlation between x and y.

(*iv*) If $r = 0$, then there is no correlation between x and y.

(*v*) If $-1 \le r < 0$, then there is negative correlation between x and y.

(*vi*) If $0 < r \le 1$, then there is positive correlation between x and y.

6.9. SPEARMAN'S RANK CORRELATION

Sometimes we have to deal with problems in which data cannot be measured quantitatively but qualitative assessment is possible, *e.g.* beauty, honesty, morality

NOTES

etc. In such a cases we assign ranks to the individuals possessing these attributes or characteristics. The best individual is given rank 1, the next rank 2 and so on.

The coefficient of rank correlation r is given by

$$r(x, y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_i is the difference of corresponding rank and n is the number of pairs of observations.

Let (x_i, y_i) ; $i = 1, 2, \ldots\ldots, n$ be the ranks of the ith individuals in two characteristics x and y respectively. Assuming that no two individuals are equal in either classification, each individual takes the values 1, 2, 3, ..., n.

Then $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(1 + 2 + 3 + \ldots\ldots + n) = \frac{1}{n}\cdot\frac{n(n+1)}{2} = \frac{n+1}{2}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(1 + 2 + 3 + \ldots\ldots + n) = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

$$\sigma_x^2 = \sigma_y^2 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2\right) - (\bar{y})^2 = \frac{1}{n}(1^2 + 2^2 + 3^2 + \ldots\ldots + n^2) - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{(n+1)}{12}(4n + 2 - 3n - 3) = \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12}$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (x_i - y_i)^2 = \sum_{i=1}^{n} [(x_i - \bar{x}) - (y_i - \bar{y})]^2 \qquad (\because \quad \bar{x} = \bar{y})$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\therefore \quad \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]$$

We know that $\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$, $\text{var}(y) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$

$$\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\therefore \quad \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\,\text{var}(x) - 2r(x, y)\,\sigma_x\,\sigma_y \qquad \left[\because \quad r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}\right]$$

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\sigma_x^2 - 2r(x, y)\,\sigma_x^2 \qquad [\because \quad \sigma_x = \sigma_y]$$

NOTES

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\sigma_x^2\,[1 - r(x, y)]$$

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\left(\frac{n^2-1}{12}\right)[1 - r(x, y)]$$

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{n^2-1}{6}\,[1 - r\,(x, y)]$$

$$\frac{6}{n\,(n^2-1)}\sum_{i=1}^{n} d_i^2 = 1 - r\,(x, y)$$

$$\Rightarrow \qquad r(x, y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\,(n^2-1)}.$$

SOLVED EXAMPLES

Example 1. *Find the coefficient of correlation for the following data :*

$n = 10, \Sigma x = 50, \Sigma y = -30, \Sigma x^2 = 290, \Sigma y^2 = 300, \Sigma xy = -115.$

Solution.
$$r(x, y) = \frac{n\,\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{10 \times (-115) - (50)\,(-30)}{\sqrt{10 \times 290 - (50)^2}\,\sqrt{10 \times 300 - (-30)^2}}$$

$$= \frac{-1150 + 1500}{\sqrt{400}\,\sqrt{2100}} = \frac{350}{200\,\sqrt{21}} = 0.3819.$$

Example 2. *A computer while calculating correlation coefficient between two variables x and y from 25 pairs of observations obtained the following results :*

$n = 25, \Sigma x = 125, \Sigma y = 100, \Sigma x^2 = 650, \Sigma y^2 = 460, \Sigma xy = 508.$

It was, however later, discovered at the time of checking that he had copied down two pairs as :

x	y
6	14
8	6

while the correct values were

x	y
8	12
6	8

. Obtain the correct value of correlation coefficient.

Solution. Corrected Σx = given Σx – (sum of incorrect values) + (sum of the correct values)

Corrected $\Sigma x = 125 - (6 + 8) + (8 + 6) = 125$

Corrected $\Sigma y = 100 - (14 + 6) + (12 + 8) = 100$

Corrected $\Sigma x^2 = 650 - (6^2 + 8^2) + (8^2 + 6^2) = 650$

Corrected $\Sigma y^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 436$

Corrected $\Sigma xy = 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8) = 520$

$$\text{Corrected } r\,(x, y) = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\,\sqrt{25 \times 436 - (100)^2}}$$

$$= \frac{500}{25 \times 30} = 0.66.$$

NOTES

Example 3. *Calculate the Karl Pearson's coefficient of correlation for the following data :*

x	*2*	*4*	*6*	*8*	*10*
y	*20*	*12*	*18*	*10*	*40*

Solution.

x	y	x^2	y^2	xy
2	20	4	400	40
4	12	16	144	48
6	18	36	324	108
8	10	64	100	80
10	40	100	1600	400
$\Sigma x = 30$	$\Sigma y = 100$	$\Sigma xy = 220$	$\Sigma y^2 = 2568$	$\Sigma xy = 676$

Here, $n = 5$

$$r(x, y) = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{5 \times 676 - 30 \times 100}{\sqrt{5 \times 220 - (30)^2}\,\sqrt{5 \times 2568 - (100)^2}}$$

$$= \frac{3380 - 3000}{\sqrt{1100 - 900}\,\sqrt{12840 - 10000}} = \frac{380}{\sqrt{200}\,\sqrt{2840}} = 0.5042$$

Example 4. *Find the Karl Pearson's coefficient of correlation between x and y for the following data :*

x	*150*	*153*	*154*	*155*	*157*	*160*	*163*	*164*
y	*65*	*66*	*67*	*70*	*68*	*53*	*70*	*63*

Solution. Let $u_i = x_i - 155$ and $v_i = y_i - 68$

x	y	u	v	u^2	v^2	uv
150	65	– 5	– 3	25	9	15
153	66	– 2	– 2	4	4	4
154	67	– 1	– 1	1	1	1
155	70	0	2	0	4	0
157	68	2	0	4	0	0
160	53	5	– 15	25	225	– 75
163	70	8	2	64	4	16
164	63	9	– 5	81	25	– 45
Total		16	– 22	204	272	– 84

Here, $n = 8$

$$r(x, y) = \frac{n\,\Sigma u_i v_i - \Sigma u_i\,\Sigma v_i}{\sqrt{n\Sigma u_i^2 - (\Sigma u_i)^2}\,\sqrt{n\Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{8 \times (-84) - (16) \times (-22)}{\sqrt{8 \times 204 - (16)^2}\,\sqrt{8 \times 272 - (-22)^2}}$$

$$= \frac{-672 + 352}{\sqrt{1376}\,\sqrt{1692}} = \frac{-320}{1525.8414} = -0.2097.$$

NOTES

Example 5. *Calculate the rank correlation coefficient for the following data :*

Student	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Rank in Maths.	*9*	*10*	*6*	*5*	*7*	*2*	*4*	*8*	*1*	*3*
Rank in Stats.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

Solution. Here, the ranks are given and $n = 10$.

Student	R_1	R_2	$d = R_1 - R_2$	d^2
A	9	1	8	64
B	10	2	8	64
C	6	3	3	9
D	5	4	1	1
E	7	5	2	4
F	2	6	– 4	16
G	4	7	– 3	9
H	8	8	0	0
I	1	9	– 8	64
J	3	10	– 7	49
				$\Sigma d^2 = 280$

$$r = 1 - \frac{6\,\Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 280}{10\,(10^2 - 1)} = 1 - \frac{1680}{990} = 1 - 1.697 = -0.697.$$

Example 6. *Ten competitors in a beauty contest are ranked by three judges in the following order :*

First Judge	*1*	*6*	*5*	*10*	*3*	*2*	*4*	*9*	*7*	*8*
Second Judge	*3*	*5*	*8*	*4*	*7*	*10*	*2*	*1*	*6*	*9*
Third Judge	*6*	*4*	*9*	*8*	*1*	*2*	*3*	*10*	*5*	*7*

Using the rank correlation method, discuss which pair of judges has the nearest approach to common taste in beauty ?

NOTES

Solution. Let R_1, R_2, R_3 be the ranks given by three judges.

R_1	R_2	R_3	$d_{12} = R_1 - R_2$	$d_{13} = R_1 - R_3$	$d_{23} = R_2 - R_3$	d_{12}^2	d_{13}^2	d_{23}^2
1	3	6	– 2	– 5	– 3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	– 3	– 4	– 1	9	16	1
10	4	8	6	2	– 4	36	4	16
3	7	1	– 4	2	6	16	4	36
2	10	2	– 8	0	8	64	0	64
4	2	3	2	1	– 1	4	1	1
9	1	10	8	– 1	– 9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	– 1	1	2	1	1	4
					Total	200	60	214

Here, $n = 10$

Rank correlation coefficient between first and second judges

$$r_{12} = 1 - \frac{6\Sigma d_{12}^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -0.212$$

Rank correlation coefficient between first and third judges,

$$r_{13} = 1 - \frac{6\Sigma d_{13}^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = 0.636$$

Rank correlation coefficient between second and third judges,

$$r_{23} = 1 - \frac{6\Sigma d_{23}^2}{n(n^2 - 1)} = 1 - \frac{6 \times 214}{10 \times 99} = 1 - \frac{214}{165} = -0.297$$

Since r_{13} is a maximum, therefore, the pair of judges first and third has the nearest approach to common tastes in beauty.

Example 7. *The marks obtained by 9 students in Statistics and Mathematics are given below :*

Marks in Statistics	*35*	*23*	*47*	*17*	*10*	*43*	*9*	*6*	*28*
Marks in Mathematics	*30*	*33*	*45*	*23*	*8*	*49*	*12*	*4*	*31*

Compute the rank correlation coefficient.

Solution. Here, the marks are given. First find the ranks and then differences.

Marks in Statistics (X)	*Marks in Mathematics (Y)*	*Ranks in X (x_i)*	*Rank in Y (y_i)*	$d_i = x_i - y_i$	d_i^2
35	30	3	5	– 2	4
23	33	5	3	2	4
47	45	1	2	– 1	1
17	23	6	6	0	0
10	8	7	8	– 1	1
43	49	2	1	1	1
9	12	8	7	1	1
6	4	9	9	0	0
28	31	4	4	0	0
				Total	12

Here, $n = 9$, $\Sigma d_i^2 = 12$

$$r = 1 - \frac{6 \times \Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 12}{9(9^2 - 1)}$$

$$= 1 - \frac{6 \times 12}{9 \times 80} = 1 - 0.1 = 0.9.$$

NOTES

EXERCISE 6.2

1. Calculate the Karl Pearson's correlation coefficient between height of father and height of son from the given data :

Height of father (in inches)	64	65	66	67	68	69	70
Height of son (in inches)	66	67	65	68	70	68	72

2. Calculate the Karl Pearson's correlation coefficient from the following data :

Overheads in ('000 ₹)	80	90	100	110	120	130	140	150	160
Cost ('000 ₹)	15	19	16	19	17	18	16	18	15

3. Calculate the Karl Pearson's correlation coefficient from the following data using 20 as the working mean for price and 70 as the working mean for demand.

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

4. Coefficient of correlation between x and y for 20 items is 0.3. Mean of x is 15 and mean of y is 20 while standard deviations are 4 and 5 for x and y respectively. At the time of calculation one item 27 has wrongly been taken as 17 in case of x series and 35 instead of 30 in case of y series. Find the correct coefficient of correlation.

5. Ten students got the following marks in Statistics and Mathematics :

Marks in Statistics	78	36	98	25	75	82	90	62	65	39
Marks in Mathematics	84	51	91	60	68	62	86	58	53	47

Calculate the coefficient of correlation.

6. Calculate the correlation coefficient from the following results :

$$n = 10,\ \Sigma x = 140,\ \Sigma y = 150,\ \Sigma(x - 10)^2 = 180,$$

$$\Sigma(y - 15)^2 = 215 \quad \text{and} \quad \Sigma(x - 10)\,(y - 15) = 60.$$

7. Calculate the coefficient of rank correlation from the following data :

x	10	12	8	15	20	25	40
y	15	10	6	25	16	12	18

8. Calculate the coefficient of rank correlation from the following data :

x	4	6	8	10	12	14	16	18
y	10	15	20	25	30	35	40	45

NOTES

9. In a beauty contest two judges rank the ten competitors in the following order :

Competitors	A	B	C	D	E	F	G	H	I	J
Rank by Judge I	6	4	3	1	2	7	9	8	10	5
Rank by Judge II	4	1	6	7	5	8	10	9	3	2

Determine if the two judges have the same taste in beauty.

10. The coefficient of rank correlation between marks in Statistics and marks in Mathematics obtained by a certain group of students is 0.8. If the sum of the squares of the differences in marks is 33, find the number of students in the group.

11. The coefficient of rank correlation of the marks obtained by 10 students in Mathematics and Statistics was found to be 0.5. It was then detected that the difference in ranks in the two subjects for one particular student was wrongly taken to be 3 in place of 7. What should be the correct rank correlation coefficients ?

Answers

1. $r = 0.81$ **2.** $r = -0.11547$ **3.** $r = -0.9542$

4. Correct $r = 0.515$ **5.** $r = 0.78$ **6.** 0.9151 **7.** $r = 0.57$

8. $r = 1$ **9.** Yes **10.** $n = 10$ **11.** $r = 0.2576$.

NOTES

7. REGRESSION ANALYSIS

STRUCTURE

Linear Regression
Lines of Regression
Properties of Regression Coefficients
Angle Between Two Lines of Regression
Standard Error of Estimate (or Prediction)
Coefficient of Determination
Properties of Coefficient of Determination

Regression analysis attempts to establish the nature of relationship between the variables. It also helps to determine the functional relationship between the variables so that one can predict or estimate the value of one variable for the given value of the other variable. Regression measures the nature and extent of correlation.

7.1. LINEAR REGRESSION

If the variables in a bivariate distribution are correlated, then points in scatter diagram will be more or less concentrated round a curve. This curve is called the curve of regression. If the curve is straight line, it is called a line of regression and the regression is said to be linear. Since the line of regression gives the best estimate to the value of dependent variable for any given value of the independent variable, therefore, it is called the line of best fit which is obtained by the method of least squares. Since any one of the two variables x and y can be taken as the independent variable and the other as a dependent variable. Therefore, there are two regression lines, one as the line of regression of y on x and the other as the line of regression of x on y.

7.2. LINES OF REGRESSION

Let the equation of line of regression of y on x be

$$y = a + bx \quad ...(1) \quad \text{then} \quad \bar{y} = a + b\bar{x} \quad ...(2)$$

NOTES

Now subtracting (2) from (1), we have

$$y - \bar{y} = b(x - \bar{x}) \quad ...(3)$$

The normal equation for the equation (1) are

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad ...(4)$$

Shifting the origin to $(\bar{x}, \bar{y})$, (4) becomes

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\,\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2 \quad ...(5)$$

We know that

$$r = \frac{\frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y})}{\sigma_x\,\sigma_y};\ \Sigma(x - \bar{x}) = 0;$$

and

$$\sigma_x^2 = \frac{1}{n}\Sigma(x - \bar{x})^2$$

From (5), we have

$$nr\,\sigma_x\sigma_y = a.0 + b.n\sigma_x^2 \quad \text{or} \quad b = r\,\frac{\sigma_y}{\sigma_x}$$

Hence, from (3), the line of regression of y on x is given by

$$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Similarly, the line of regression of x on y is given by

$$x - \bar{x} = r\,\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$r\,\frac{\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x and is denoted by b_{yx}.

$$b_{yx} = \frac{r\,\sigma_y}{\sigma_x} = \frac{\text{cov}(x, y)}{\sigma_x^2} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$r\,\frac{\sigma_x}{\sigma_y}$ is called the regression coefficient of x on y and is denoted by b_{xy}.

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y} = \frac{\text{cov}(x, y)}{\sigma_y^2} = \frac{n\,\Sigma xy - \Sigma x\Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2}$$

Note. The line of regression of y on x is used to estimate the value of y for given value of x. The line of regression of x on y is used to estimate the value of x for given value of y.

7.3. PROPERTIES OF REGRESSION COEFFICIENTS

(*i*) The correlation coefficient and two regression coefficients are of the same sign.

(*ii*) If one of the regression coefficient is greater than unity, the other must be less than unity.

(*iii*) Arithmetic mean of regression coefficients is greater than the correlation coefficient.

(*iv*) The correlation coefficient is the geometric mean between the regression coefficients.

(*v*) Regression coefficients are independent of the origin and not of scale.

NOTES

7.4. ANGLE BETWEEN TWO LINES OF REGRESSION

If θ is the acute angle between the two lines of regression in the case of two variables x and y, then

$$\tan\theta = \frac{1-r^2}{r^2} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2},$$

where r, σ_x and σ_y have their usual meanings.

Explain the significance when $r = 0$ and $r = \pm 1$.

Proof. Equation of the line of regression y on x is

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

and the equation of the line of regression x on y is

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

Their slopes are $m_1 = r \frac{\sigma_y}{\sigma_x}$ and $m_2 = \frac{\sigma_y}{r\,\sigma_x}$

$$\therefore \quad \tan\theta = \pm \frac{m_2 - m_1}{1 + m_1 m_2} = \pm \frac{\frac{\sigma_y}{r\,\sigma_x} - r\frac{\sigma_y}{\sigma_x}}{1 + \frac{\sigma_y^2}{\sigma_x^2}}$$

$$= \pm \frac{1-r^2}{r} \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} = \pm \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

Since $r^2 \le 1$ and σ_x, σ_y are positive.

$\therefore$ Positive sign gives the acute angle between the lines.

Hence, $$\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

When $r = 0$, $\theta = \frac{\pi}{2}$.

So two lines of regression are perpendicular to each other.

When $r = \pm 1$, $\tan\theta = 0$ so that $\theta = 0$ or π.

So two lines of regression coincide and there is perfect correlation between the two variables x and y.

SOLVED EXAMPLES

Example 1. *Find the equation of two lines of regression for the data :*

x	*1*	*2*	*3*	*4*	*5*
y	*7*	*6*	*5*	*4*	*3*

and hence find an estimate of y for x = 3.5 from the appropriate line of regression.

NOTES

Solution.

x	y	x^2	y^2	xy
1	7	1	49	7
2	6	4	36	12
3	5	9	25	15
4	4	16	16	16
5	3	25	9	15
$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma x^2 = 55$	$\Sigma y^2 = 135$	$\Sigma xy = 65$

Here, $n = 5$

$$\bar{x} = \frac{1}{n}\Sigma x_i = \frac{15}{5} = 3, \ \bar{y} = \frac{1}{n}\Sigma y_i = \frac{25}{5} = 5$$

Now,
$$b_{yx} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{5\times 65 - 15\times 25}{5\times 55 - (15)^2} = \frac{13-15}{11-9} = -1$$

$$b_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma y_i^2 - (\Sigma y_i)^2} = \frac{5\times 65 - 15\times 25}{5\times 135 - (25)^2} = \frac{13-15}{27-25} = -1$$

So, line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad \Rightarrow \quad y - 5 = -1\,(x-3)$$

$$\Rightarrow \qquad y = -x + 8$$

and the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}) \quad \Rightarrow \quad x - 3 = -1\,(y-5)$$

$$\Rightarrow \qquad x = -y + 8$$

To estimate the value of y when x is given, we use the line of regression of y on x, *i.e.*

$$y = -x + 8$$

Now substitute $x = 3.5$, we have

$$y = -3.5 + 8 = 4.5.$$

Example 2. *The following table gives age (x) in years of cars and annual maintenance cost (y) in hundred rupees :*

x	*1*	*3*	*5*	*7*	*9*
y	*15*	*18*	*21*	*23*	*22*

Estimate the maintenance cost for a 6 years old car after finding the appropriate line of regression.

Solution.

x	y	x^2	xy
1	15	1	15
3	18	9	54
5	21	25	105
7	23	49	161
9	22	81	198
$\Sigma x = 25$	$\Sigma y = 99$	$\Sigma x^2 = 165$	$\Sigma xy = 533$

Here, $n = 5$

$$\bar{x} = \frac{1}{n}\Sigma x_i = \frac{25}{5} = 5,\ \bar{y} = \frac{1}{n}\Sigma y_i = \frac{99}{5} = 19.8$$

$$b_{yx} = \frac{n\,\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\,\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{5 \times 533 - 25 \times 99}{5 \times 165 - (25)^2}$$

$$= \frac{2665 - 2475}{825 - 625} = \frac{190}{200} = 0.95$$

Regression line of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 19.8 = 0.95\,(x - 5)$$

$$y = 0.95\,x + 15.05$$

When $x = 4$ years.

$$y = 0.95 \times 4 + 15.05$$

$$= 18.85 \text{ hundred rupees}$$

$$= ₹\ 1885.$$

Example 3. *From the following information on values of two variables x and y, find the two regression lines and the correlation coefficient between x and y.*

$$n = 10,\ \Sigma x = 20,\ \Sigma y = 40,\ \Sigma x^2 = 240,$$

$$\Sigma y^2 = 410,\ \Sigma xy = 200.$$

Solution. We know that

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x \Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{10 \times 200 - 20 \times 40}{10 \times 240 - (20)^2} = \frac{20 - 8}{24 - 4} = \frac{12}{20} = \frac{3}{5}$$

and

$$b_{xy} = \frac{n\,\Sigma xy - \Sigma x \Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{10 \times 200 - 20 \times 40}{10 \times 410 - (40)^2} = \frac{20 - 8}{41 - 16} = \frac{12}{25}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{20}{10} = 2,\ \bar{y} = \frac{\Sigma y}{n} = \frac{40}{10} = 4$$

The two regression lines are

$$y - \bar{y} = b_{yx}(x - \bar{x}),\quad i.e.\ y - 4 = \frac{3}{5}(x - 2)$$

$$y = 0.6\,x + 2.8$$

and

$$x - \bar{x} = b_{xy}(y - \bar{y})\ i.e.,\quad x - 2 = \frac{12}{25}(y - 4)$$

$$x = 0.48\,y + 0.08$$

We know that $r = \pm\sqrt{b_{yx} \times b_{xy}}$

$$r = \sqrt{\frac{3}{5} \times \frac{12}{25}} = \sqrt{\frac{36}{125}} = 0.536$$

($\because$ Regression coefficients are positive so r will be positive)

Example 4. *For 100 students of a class, the regression equation of marks in Statistics (x) on the marks in Mathematics (y) is 3y – 5x + 180 = 0. The mean marks in Mathematics is 50 and variance of marks in Statistics is* $\frac{4}{9}$ *th of the variance of marks*

NOTES

NOTES

in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between marks in the two subjects.

Solution. Since the given line of regression is x on y so we have

$$x = \frac{3}{5}y + \frac{180}{5} = \frac{3}{5}y + 36$$

We have $b_{xy} = \frac{3}{5} = r\frac{\sigma_x}{\sigma_y}$

Given variance of $x = \frac{4}{9}$ variance of y

$$\Rightarrow \quad \frac{\text{Variance of } x\ (\sigma_x{}^2)}{\text{Variance of } y\ (\sigma_y{}^2)} = \frac{4}{9} \qquad \Rightarrow \quad \frac{\sigma_x}{\sigma_y} = \frac{2}{3}$$

$$\therefore \qquad \frac{3}{5} = r \times \frac{2}{3} \qquad \Rightarrow \qquad r = \frac{9}{10} = 0.9$$

($\because$ b_{xy} is positive, r is positive)

Since the mean of x and mean of y lie on the regression lines, we have

$$\bar{x} = \frac{3}{5}\bar{y} + 36 \quad \Rightarrow \bar{x} = \frac{3}{5} \times 50 + 36 = 66. \quad (\because \quad \bar{y} = 50)$$

Example 5. *The lines of regression of y on x and x on y are $y = x + 5$ and $16x - 9y = 94$ respectively.*

Find the variance of x if the variance of y is 16. Also find the covariance of x and y.

Solution. Regression equation of y on x is

$$y = x + 5 \quad \Rightarrow \quad b_{yx} = 1 \qquad \text{(Coefficient of } x\text{)}$$

Regression equation of x on y is

$$16x - 9y = 94, \quad i.e., \quad x = \frac{9}{16}y + \frac{94}{16}$$

$$\Rightarrow \qquad b_{xy} = \frac{9}{16} \qquad \text{(Coefficient of } y\text{)}$$

We know that $r = \pm\sqrt{b_{yx} \times b_{xy}}$

$$r = \sqrt{1 \times \frac{9}{16}} = \frac{3}{4} = 0.75 \quad (r \text{ is positive since } b_{yx} \text{ and } b_{xy} \text{ are positive})$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad \Rightarrow \quad \sigma_x = \frac{b_{xy} \times \sigma_y}{r} = \frac{\left(\frac{9}{16}\right) \times (4)}{\left(\frac{3}{4}\right)} = 3 \ (\because \quad \sigma_y{}^2 = 16)$$

$$r = \frac{\text{Cov.}\,(x, y)}{\sigma_x\ \sigma_y}$$

$$\Rightarrow \qquad \text{cov}\ (x, y) = r\ \sigma_x\ \sigma_y = \frac{3}{4} \times 3 \times 4 = 9.$$

Example 6. *From the following information on values of two variables x and y, find the two regression lines and estimate values of x and y if $y = 10$ and $x = 8$ respectively.*

$$n = 5,\ \Sigma x = 15,\ \Sigma y = 18,\ \Sigma x^2 = 55,\ \Sigma y^2 = 74,\ \Sigma xy = 58.$$

Solution. We know that

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5\times 58 - 15\times 18}{5\times 55 - (15)^2} = \frac{290-270}{275-225} = \frac{20}{50} = \frac{2}{5}$$

and

$$b_{xy} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{5\times 58 - 15\times 18}{5\times 74 - (18)^2} = \frac{290-270}{370-324} = \frac{20}{46} = \frac{10}{23}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3 \quad \text{and} \quad \bar{y} = \frac{\Sigma y}{n} = \frac{18}{5}$$

The line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - \frac{18}{5} = \frac{2}{5}(x - 3)$$

$$y = 0.4x + 2.4 \qquad ...(1)$$

The line of regression of x on y is given by

$$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 3 = \frac{10}{23}\left(y - \frac{18}{5}\right) \text{ or } x = 0.435y + 1.435 \qquad ...(2)$$

Putting $x = 8$ in (1), we have

$$y = 0.4 \times 8 + 2.4 = 3.2 + 2.4 = 5.6$$

Putting $y = 10$ in (2), we have

$$x = 0.435 \times 10 + 1.435 = 4.35 + 1.435 = 5.785.$$

Example 7. *The information about advertising and sales of a manufacturing concern is given as follows :*

	Advertising expenditure (*x*) (₹ *Lacs*)	*Sales* (*y*) (₹ *Lacs*)
Mean	*10*	*90*
S.D.	*3*	*12*

Correlation coefficient = 0.8.

Find (i) the regression coefficients b_{yx} and b_{xy}.

(*ii*) *the two regression lines*

(*iii*) *the likely sales when advertisement expenditure is ₹ 18 lacs.*

(*iv*) *the advertisement expenditure if the company wants to attain sales target of ₹ 115 lacs.*

Solution. Given $\bar{x} = 10$, $\bar{y} = 90$, $\sigma_x = 3$, $\sigma_y = 12$ and $r = 0.8$

(*i*)
$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.8 \times \frac{12}{3} = 3.2$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.8 \times \frac{3}{12} = 0.2$$

(*ii*) The line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 90 = 3.2(x - 10)$$

$$y = 3.2x + 58$$

NOTES

The line of regression of x on y is given by

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 10 = 0.2(y - 90)$$

$$x = 0.2y - 8$$

(*iii*) Given $x = 18$ we have to estimate y

$\therefore$ $y = 3.2 \times 18 + 58 =$ Rs. 115.6 lacs

(*iv*) Given $y = 115$ we have to estimate x

$\therefore$ $x = 0.2 \times 115 - 8 =$ Rs. 15 lacs.

Example 8. *The equations of two lines of regression are*

$$4x + 3y + 7 = 0 \quad \text{and} \quad 3x + 4y + 8 = 0.$$

Find (*i*) *mean value of x and y*

(*ii*) *the regression coefficients* b_{yx} *and* b_{xy}

(*iii*) *the correlation coefficient between x and y.*

(*iv*) *the standard deviation of y, if the variance of x is 4*

(*v*) *the value of y for x = 5*

Solution. (*i*) Since the mean of x and mean of y lie on the regression lines, we have

$\therefore$ $4\bar{x} + 3\bar{y} + 7 = 0$ or $4\bar{x} + 3\bar{y} = -7$

and $3\bar{x} + 4\bar{y} + 8 = 0$ or $3\bar{x} + 4\bar{y} = -8$

On solving the above equations for $\bar{x}$ and $\bar{y}$, we have

$$\bar{x} = -\frac{4}{7} \quad \text{and} \quad \bar{y} = -\frac{11}{7}$$

Mean of $x = -\frac{4}{7}$ and mean of $y = -\frac{11}{7}$

(*ii*) Let the regression line of y on x be

$$4x + 3y + 7 = 0 \quad \text{or} \quad y = -\frac{4}{3}x - \frac{7}{3}$$

$\therefore$ $b_{yx} = -\frac{4}{3}$

and the regression line of x on y be

$$3x + 4y + 8 = 0 \quad \text{or} \quad x = -\frac{4}{3}y - \frac{8}{3}$$

$\therefore$ $b_{xy} = -\frac{4}{3}$

Since $b_{yx} \cdot b_{xy} = \left(-\frac{3}{4}\right)\left(-\frac{3}{4}\right) = \frac{9}{16} < 1$

Hence, the choice of regression line is correct.

So $b_{yx} = -\frac{3}{4}$ and $b_{xy} = -\frac{3}{4}$

(*iii*) We know that $r = \pm\sqrt{b_{yx} \times b_{xy}}$

$\therefore$ $r = \pm\sqrt{\left(-\frac{3}{4}\right) \times \left(-\frac{3}{4}\right)} = \pm\frac{3}{4} = -\frac{3}{4}$

($\because$ b_{yx} and b_{xy} have the negative sign)

(*iv*) We have $\sigma_x^2 = 4 \Rightarrow \sigma_x = 2$

Now, $b_{yx} = -\frac{3}{4}$ or $r\frac{\sigma_y}{\sigma_x} = -\frac{3}{4}$

$$\left(-\frac{3}{4}\right) \times \frac{\sigma_y}{2} = -\frac{3}{4} \Rightarrow \sigma_y = 2$$

(*v*) Since we have to find y when x is given, we use line of regression of y on x

$$y = -\frac{4}{3}x - \frac{7}{3}$$

Putting $x = 5$, we have

$$y = -\frac{4}{3} \times 5 - \frac{7}{3} = -6.67 - 2.33$$

$$y = -9.$$

Example 9. *Consider the two regression lines :*

$$3x + 2y = 26 \quad \text{and} \quad 6x + y = 31.$$

(*i*) *Find the mean values of x and y.*

(*ii*) *Find the correlation coefficient between x and y.*

(*iii*) *Show that the estimated value of y for x = 0 is 13 whereas estimated value of x for y = 13 is 3.*

Solution. (*i*) Since the mean of x and mean of y lie on the regression line, we have

$$3\bar{x} + 2\bar{y} = 26 \quad \text{and} \quad 6\bar{x} + \bar{y} = 31$$

On solving these two equations for $\bar{x}$ and $\bar{y}$,

We have $\bar{x} = 4$ and $\bar{y} = 7$

$\therefore$ Mean of $x = 4$ and mean of $y = 7$

(*ii*) Let the regression line of y on x be

$$3x + 2y = 26 \quad \text{or} \quad y = -\frac{3}{2}x + 13 \quad \therefore \quad b_{yx} = -\frac{3}{2}$$

and let the regression line of x on y be

$$6x + y = 31 \quad \text{or} \quad x = -\frac{1}{6}y + \frac{31}{6} \quad \therefore \quad b_{xy} = -\frac{1}{6}$$

We know that $r = \pm\sqrt{b_{yx} \times b_{xy}}$

$$r = \pm\sqrt{-\frac{3}{2} \times -\frac{1}{6}} = \pm\frac{1}{2} = -\frac{1}{2}$$

(r is negative since b_{yx} and b_{xy} are negative)

(*iii*) Since we have to show the estimated value of y for the given value of x, we use line of regression of y on x

$$y = -\frac{3}{2}x + 13$$

Putting $x = 0$, we have $y = -\frac{3}{2} \times 0 + 13 = 13$

To show the estimated value of x for the given value of y, we use line of regression of x on y

$$x = -\frac{1}{6}y + \frac{31}{6}$$

NOTES

Putting $y = 13$, we have

$$x = -\frac{1}{6} \times 13 + \frac{31}{6} = \frac{18}{6} = 3.$$

NOTES

7.5. STANDARD ERROR OF ESTIMATE (OR PREDICTION)

The square root of arithmetic mean of squared deviation of the predicted value from the observed value is known as the standard error of estimate or prediction. It is given by

$$E_{yx} = \sqrt{\frac{\Sigma(y - y_p)^2}{n}},$$

where y is the actual value and y_p is the predicted value ; E_{yx} is called the standard error of estimate or prediction of y on x.

Example. *Find the standard error of estimate of y on x from the following data :*

x	*1*	*2*	*3*	*4*	*5*
y	*2*	*5*	*3*	*8*	*7*

Solution. For the standard error of estimate of y on x we have to find regression line of y on x.

x	y	x^2	xy
1	2	1	2
2	5	4	10
3	3	9	9
4	8	16	32
5	7	25	35
$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma x^2 = 55$	$\Sigma xy = 88$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{25}{5} = 5$$

Here, $n = 5$

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x \Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{5 \times 88 - 15 \times 25}{5 \times 55 - (15)^2} = \frac{65}{50} = 1.3$$

The line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 5 = 1.3\,(x - 3)$$

$$y = 1.3x + 1.1 \quad \text{or} \quad y_p = 1.3x + 1.1$$

Now

x	y	$y_p = 1.3x + 1.1$	$y - y_p$	$(y - y_p)^2$
1	2	$1.3 \times 1 + 1.1 = 2.4$	– 0.4	0.16
2	5	$1.3 \times 2 + 1.1 = 3.7$	1.3	1.69
3	3	$1.3 \times 3 + 1.1 = 5.0$	– 2.0	4.00
4	8	$1.3 \times 4 + 1.1 = 6.3$	1.7	2.89
5	7	$1.3 \times 5 + 1.1 = 7.6$	– 0.6	0.36
				$\Sigma(y - y_p)^2 = 9.10$

$$E_{yx} = \sqrt{\frac{\Sigma(y - y_p)^2}{n}} = \sqrt{\frac{9.10}{5}} = \sqrt{1.82} = 1.349.$$

7.6. COEFFICIENT OF DETERMINATION

The quantity r^2 is called the coefficient of determination. It lies between 0 and 1.

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\Sigma(y_p - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

The quantity $(1 - r^2)$ is called the coefficient of non-determination. Also the quantity $\sqrt{1 - r^2}$ is called the coefficient of alienation.

Since r lies between – 1 and + 1, r^2 lies between 0 and 1 both inclusive.

Note. This is another formula to calculate correlation coefficient r.

7.8. PROPERTIES OF COEFFICIENT OF DETERMINATION

(*i*) As an index of fit it is interpreted as the total proportion of variance in y explained by x.

(*ii*) As a measure of linear relationship it tells us how well the regression line fits the data.

(*iii*) As an important indicator of the predictive accuracy of the regression equation, the minimum value of r^2 should be 0.8, otherwise, the predictive accuracy is considered low.

EXERCISE 7.1

1. Find the line of regression of y on x for the following data :

x	10	9	8	7	6	4	3
y	8	12	7	10	8	9	6

NOTES

2. Find the line of regression of y on x for the following data :

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Estimate the value of y, when $x = 10$.

3. Find the regression lines for the following data :

x	6	2	10	4	8
y	9	11	5	8	7

4. Find the regression coefficient b_{xy} between x and y for the following data :
$\Sigma x = 30$, $\Sigma y = 42$, $\Sigma xy = 199$, $\Sigma x^2 = 184$, $\Sigma y^2 = 318$ and $n = 6$

5. Find the regression coefficients b_{yx} and b_{xy} for the following data :
$\Sigma x = 24$, $\Sigma y = 12$, $\Sigma x^2 = 374$, $\Sigma y^2 = 97$, $\Sigma xy = 157$ and $n = 7$.
Also, find the coefficient of correlation between x and y.

6. Find the regression line of x on y and estimate the value of x, when $y = 5$ from the following data :
$\Sigma x = 125$, $\Sigma y = 100$, $\Sigma x^2 = 1650$, $\Sigma y^2 = 1500$, $\Sigma xy = 50$ and $n = 25$.

7. The following regression equations were obtained from a correlation table :
$y = 0.516\,x + 33.73$; $x = 0.512\,y + 32.52$
Find the value of
(*i*) the mean of *x's* and the mean of *y's* (*ii*) the correlation coefficient.
(*iii*) the coefficient of determination.

8. You are given the following data :

Series	x	y
Mean	18	100
Standard deviation	14	20

Correlation coefficient between x and $y = 0.8$.
Find (*i*) the regression coefficients b_{yx} and b_{xy}.
(*ii*) the two regression lines.
(*iii*) estimate the value of y, when $x = 70$.
(*iv*) estimate the value of x, when $y = 90$.

9. If $4x - 5y + 33 = 0$ and $20x - 9y - 107 = 0$ are two lines of regression. Find
(*i*) the mean values of x and y.
(*ii*) the regression coefficients b_{yx} and b_{xy}.
(*iii*) the correlation coefficient between x and y.
(*iv*) the standard deviation of y, if variance of x is 9.
(*v*) the coefficient of determination.

10. Find the standard error of estimate of y on x for the following data :

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

11. In a partially destroyed record, for the estimation of the two lines of regression from a bivariate data (x, y), the following results were available :
Regression coefficient of y on $x = -1.6$, regression coefficient of x on $y = -0.4$, standard error of the estimate of y on $x = 3$.
Find (*i*) coeff. of correlation between x and y (*ii*) standard deviation σ_x and σ_y (*iii*) standard error of estimate of x on y.

Answers

1. $\left[y = \frac{1}{3} x + \frac{133}{21} \right]$ **2.** $\left[y = \frac{7}{11} x + \frac{6}{11}, 6.91 \right]$ **3.** $y = 11.9 - 0.65\ x$; $x = 16.4 - 1.3\ y$

4. – 0.46 **5.** $b_{yx} = 0.397$; $b_{xy} = 1.516$; $r = 0.776$

6. $\left[x = -\frac{9}{22} y + \frac{146}{22} ; 4.591 \right]$ **7.** (*i*) 67.6, 68.61 (*ii*) 0.514 (*iii*) 0.264

8. (*i*) $b_{yx} = 1.14$ and $b_{xy} = 0.56$ (*ii*) $y = 1.14x + 79.41$; $x = 0.56y - 38$ (*iii*) 159.21 (*iv*) 12.4

9. (*i*) 13, 17 (*ii*) $b_{yx} = \frac{4}{5}$, $b_{xy} = \frac{9}{20}$ (*iii*) $r = 0.6$ (*iv*) $\sigma_y = 4$ (*v*) $r^2 = 0.36$

10. 0.564 **11.** (*i*) $r = -0.8$ (*ii*) $\sigma_x = 2.5$, $\sigma_y = 5$, (*iii*) $E_{xy} = 1.5$

NOTES

NOTES

8. STATISTICAL QUALITY CONTROL

STRUCTURE

8.1. INTRODUCTION

Statistical quality control (SQC) is one of the major area of production management. It is a specialised professional technique which is used to maintain the technical efficiency of the processes of production. SQC is a simple statistical method for determining the extent to which quality goals are being met without necessarily checking each and every item produced and for indicating whether or not the variations which occur are exceeding normal expectations. SQC enables us to decide whether to reject or accept a particular product.

8.2. CAUSES OF VARIATIONS

Products of exactly the same quality are not possible to be produced in the continuous flow of any manufacturing process. So the variations in quality of the product remains inevitable. These variations occurs due to two types of causes :

(*i*) **Chance or Random causes.** Some deviations from the desired specifications are bound to occur in the items produced, howsoever efficient, the production process may be. If the variations occurs due to some inherent pattern of variation and no causes can be assigned to it, it is called chance or random variation.

For instance, slight variation in temperature, pressure and humidity, etc. interact randomly to produce slight variation in the quality of the product. Chance variation is

tolerable and does not materially affect the quality of a product. In such a situation, the process is said to be under statistical control.

(*ii*) **Assignable causes.** Assignable causes (also called non random or systematic) can be easily identified. The assignable cause may occur at any stage of the process. These causes can be easily removed. Assignable causes of variation may be due to defective raw material, negligence of the operators, improper handling of machines, faulty equipments, etc. In such a situation, the process is said to be out of control.

NOTES

8.3. METHODS OF STATISTICAL QUALITY CONTROL

To control the quality characteristics of the product, there are two main methods :

1. **Process control.** The main aim in any production process is to control and maintain the quality of the product to requisite standard during the manufacturing process. This is termed as process control and is achieved through the use of control charts given by W.A. Shewhart.

2. **Product control.** This technique is concerned with inspection of already manufactured product to ascertain whether they are acceptable to the consumer or not. This is achieved through an acceptance inspection or a sampling inspection plan. Such a sampling inspection is often termed as product control.

8.4. ADVANTAGES OF STATISTICAL QUALITY CONTROL

1. SQC makes it possible to discriminate whether deviation from the requisite standard occurring in the product during manufacturing process is due to chance causes or due to assignable causes.
2. SQC is extremely useful, particularly in the case where the units are destroyed under inspection, *e.g.*, the life of an electric bulb, explosiveness of crackers, bombs, life of a battery cell, etc.
3. SQC enables to determine whether the quality standards are being met without inspecting the every unit produced.
4. SQC helps to know whether the manufacturing process is under control or not and if it has gone out of control, remedial measures can be applied.
5. SQC reduces the waste of time and material to the absolute minimum by giving an early warning about the occurrence of the defects.
6. The greatest advantage is the low cost of inspection..
7. SQC minimizes the risk of the consumer as well as the producer.
8. SQC provides protection to the manufacturer against losses due to the rejection of manufacturing products, likely to be made later on.
9. Efficient utilization of personnel, machines and materials results in higher production.
10. Removal of bottle necks in the production process.

8.5. CONTROL CHARTS

Control charts are the devices to describe the patterns of variation. The control charts were developed by W.A. Shewhart of Bell Telephone Laboratories in 1924. Based

NOTES

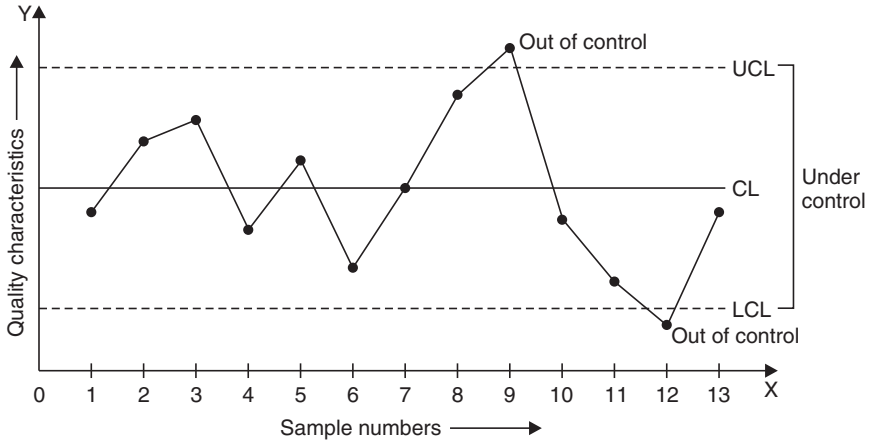
on the theory of probability and sampling, it enabled us to detect the presence of assignable causes of erratic variations in the process. These causes are then identified and eliminated and the process is stabilized and controlled at desired performances.

A control chart is the running record graph of the performance of some quality characteristics. A control chart consists of the following three horizontal lines on the graph :

(*i*) a control or central line (CL) depicting the desired standard or level of the process.

(*ii*) an upper control limit (UCL).

(*iii*) a lower control limit (LCL).



The control chart has a horizontal scale that represents the consecutive sample number and a vertical scale that represents the quality characteristic of each sample.

8.6. TYPES OF CONTROL CHARTS

Control charts are of two types depending on whether a given characteristic is measurable or not.

(*i*) **Control Charts for Variables.** These charts are used to achieve and maintain an acceptable quality level for a process whose product can be subjected to quantitative measurements.

(*ii*) **Control Charts for Attributes.** These charts are used to achieve and maintain an acceptable quality level for a process whose product cannot be subjected to quantitative measurements but can be classified as good or bad, acceptable or non-acceptable.

8.7. CONTROL CHARTS FOR VARIABLES

The most common charts for variables are

(*i*) Control Charts for sample means ($\bar{x}$-Charts)

(*ii*) Control Charts for sample ranges (R-Charts)

NOTES

The various steps to construct $\bar{x}$ and R-charts are as follows :

1. A random sample of size n (n is usually 4 or 5 units) is taken during the manufacturing process over a period of time and the quality measurements $x_1, x_2, ..., x_n$ are noted.

2. The sample mean $\bar{x}$ and sample range R are calculated by using

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and $$R = x_{max} - x_{min},$$

where x_{max} and x_{min} are the largest and smallest values of measurements $x_1, x_2, ..., x_n$ respectively.

3. If the process is found to be satisfactory, k successive samples (k usually varies from 20 to 30) are taken and for each sample, mean $\bar{x}$ and range R are calculated. Then find the combined mean $\bar{\bar{x}}$ and combined range $\bar{R}$ by using

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + ... + \bar{x}_k}{k} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i$$

and $$\bar{R} = \frac{R_1 + R_2 + ... + R_k}{k} = \frac{1}{k}\sum_{i=1}^{k} R_i$$

4. Calculation of control limits for $\bar{x}$-chart

Control or central line (CL) = $\bar{\bar{x}}$

Upper control limit (UCL) = $\bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$

or Upper control limit (UCL) = $\bar{\bar{x}} + A_2\bar{R}$

Lower control limit (LCL) = $\bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}}$

or Lower control limit (LCL) = $\bar{\bar{x}} - A_2\bar{R}$,

where d_2 and A_2 can be found from the table depending upon the size of the sample n.

5. Calculation of control limits for R-chart

Control or central line (CL) = $\bar{R}$

Upper control limit (UCL) = $D_4\bar{R}$

Lower control limit (LCL) = $D_3\bar{R}$,

where D_3 and D_4 can be found from the table depending upon the size of the sample n.

6. The natural tolerance limits (upper and lower tolerance limits) for individual values of x are calculated by using

$$UTL_{\bar{x}} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2}$$

$$LTL_{\bar{x}} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2}$$

NOTES

The process is said to be capable of meeting the customers specifications if these natural tolerance limits fall within the customers specifications.

The process capability is = 6 $\sigma = \frac{6\overline{R}}{d_2}$, where σ is standard deviation.

Table

Sample size (n)	A_2	D_3	D_4	d_2
2	1.88	0.00	3.27	1.13
3	1.02	0.00	2.57	1.69
4	0.73	0.00	2.28	2.06
5	0.58	0.00	2.11	2.33
6	0.48	0.00	2.00	2.53
7	0.42	0.08	1.92	2.70
8	0.37	0.14	1.86	2.85
9	0.34	0.18	1.82	2.97
10	0.31	0.22	1.78	3.08
11	0.29	0.26	1.74	3.17
12	0.27	0.28	1.72	3.26
13	0.25	0.31	1.69	3.34
14	0.24	0.33	1.67	3.41
15	0.22	0.35	1.65	3.47
16	0.21	0.36	1.64	3.53
17	0.20	0.38	1.62	3.59
18	0.19	0.39	1.61	3.64
19	0.19	0.40	1.60	3.69
20	0.18	0.41	1.59	3.74

8.8. CONTROL CHARTS FOR ATTRIBUTES

Sometimes it becomes impossible to determine the quality of a product by means of measurement. The product is classified as good or bad, acceptable or non-acceptable. At times, the product is inspected for defects. Such characteristics are called attributes. The most common charts for attributes are

(*i*) Control chart for fraction defective (*p*-chart)

(*ii*) Control chart for number of defective (*np*-chart)

(*iii*) Control chart for number of defects (*c*-chart).

8.9. (i) CONTROL CHART FOR FRACTION DEFECTIVES (p-CHART)

Let n be the sample size taken from the production process at different time intervals. If d be the number of defectives in this sample of size n, then the fraction defective in this sample is given by $p = \frac{d}{n}$ or $d = np$

NOTES

If $\bar{p}$ represents the average fraction defective from all the samples (k samples) inspected, then

$$\bar{p} = \frac{\text{Total number of defectives in all the samples inspected}}{\text{Total number of items inspected in all the samples}}.$$

The Binomial distribution is used to construct the 'p' chart. By Binomial distribution standard deviation (σ_p) is given by

$$\sigma_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Control limits for p-chart are given by

$$CL_p = \bar{p}$$

$$UCL_p = \bar{p} + 3\sigma_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$LCL_p = \bar{p} - 3\sigma_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Since the number of defectives (or fraction defectives) cannot be negative, if LCL comes out to be negative, it is taken as zero.

To construct the p-chart, p-values are taken on the y-axis and sample numbers on the x-axis. If any point lies outside the control limits, it is concluded that the process is not under control otherwise under control.

8.9. (ii) CONTROL CHART FOR NUMBER OF DEFECTIVES (np-CHART)

If n is the sample size and d is the number of defectives in this sample, then $d = np$, where p is the fraction defectives in the sample.

Now, let if $n\bar{p}$ represents the average number of defectives per sample of constant size, *i.e.*, $n\bar{p} = \dfrac{\text{Total number of defective items in all the samples inspected}}{\text{Number of samples inspected}}$

Now the standard deviation (σ_{np}) is given by

$$\sigma_{np} = n\sigma_p = n\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{n\bar{p}(1-\bar{p})}$$

Control limits for np-chart are given by

$$CL_{np} = n\bar{p}$$

$$UCL_{np} = n\bar{p} + 3\sigma_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$LCL_{np} = n\bar{p} - 3\sigma_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

Since the number of defectives cannot be negative, if LCL comes out to be negative, it is taken as zero. To construct the np-chart, np, *i.e.*, d values are taken on the y-axis and sample numbers on the x-axis. If any point lies outside the control limits, it is concluded that the process is not under control otherwise under control.

8.9. (iii) CONTROL CHART FOR NUMBER OF DEFECTS (c-CHART)

NOTES

There are many situations in which it will be advantageous to know the number of defects in an item or product after classifying that item or product is defective. In this situation, *c*-chart is used. Sample size for *c*-chart may be single unit like a radio, a match box, a computer, an aircraft or a group of units. The number of defects may be 0, 1, 2, ... The variate values are discrete in nature and hence, it will follow a discrete distribution. The Poisson distribution is used to construct the *c*-chart. Since for a Poisson distribution mean and variance are same, then the standard deviation (σ_c) is given by

$$\sigma_c = \sqrt{\bar{c}},$$

where $\bar{c}$ is average number of defects in a sample.

$$\bar{c} = \frac{\text{Total number of defects in all the samples inspected}}{\text{Number of samples inspected}}$$

Control limits for *c*-chart are given by

$$CL_c = \bar{c}$$

$$UCL_c = \bar{c} + 3\sigma_c = \bar{c} + 3\sqrt{\bar{c}}$$

$$LCL_c = \bar{c} - 3\sigma_c = \bar{c} - 3\sqrt{\bar{c}}$$

Since the number of defects cannot be negative, if LCL comes out to be negative, it is taken as zero. To construct the *c*-chart, *c*-values are taken on the *y*-axis and sample numbers on the *x*-axis. If any point lies outside the control limits, it is concluded that the process is not under control otherwise under control.

SOLVED EXAMPLES

Example 1. *Using the following data calculate the control limits for $\bar{x}$-chart:*

$n = 12$, $\bar{\bar{x}} = 138.6$, $\bar{R} = 7.4$ and $d_2 = 3.258$.

Solution. The control limits for $\bar{x}$-chart are calculated as :

$$CL = \bar{\bar{x}} = 138.6$$

$$UCL = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = 138.6 + \frac{3 \times 7.4}{3.258\sqrt{12}}$$

$$= 138.6 + 1.967 = 140.567$$

$$LCL = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 138.6 - \frac{3 \times 7.4}{3.258\sqrt{12}}$$

$$= 138.6 - 1.967 = 136.633.$$

Example 2. *Using the following data calculate the control limits for R-chart:*

$n = 4$, $\bar{R} = 9.60$, $d_2 = 2.059$ and $d_3 = 0.880$.

Solution. The control limits for R-chart are calculated as :

$$CL = \bar{R} = 9.60$$

$$UCL = \bar{R} + \frac{3d_3\bar{R}}{d_2} = 9.60 + \frac{3 \times 0.880 \times 9.60}{2.059}$$

$$= 9.60 + 12.3089 = 21.91$$

$$LCL = \bar{R} - \frac{3d_3\bar{R}}{d_2} = 9.60 + \frac{3 \times 0.880 \times 9.60}{2.059}$$

$$= 9.60 - 12.3089 = -2.71.$$

NOTES

Example 3. *A machine is set to deliver packets of a given weight, 10 samples of size 5 each were recorded and the mean and range of each sample is as follows :*

Sample No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Mean ($\bar{x}$)	*49*	*45*	*48*	*53*	*39*	*47*	*46*	*39*	*51*	*45*
Range (R)	*7*	*5*	*7*	*9*	*5*	*8*	*8*	*6*	*7*	*6*

Calculate the control limits of $\bar{x}$ *and R-charts. Comment on the state of control without drawing the charts.*

Solution. Here, $n = 5$, $k = 10$, $A_2 = 0.58$, $D_3 = 0$

and $D_4 = 2.11$ (from table for $n = 5$)

$$\bar{\bar{x}} = \frac{\Sigma\bar{x}}{k} = \frac{49 + 45 + \ldots + 45}{10} = \frac{462}{10} = 46.2$$

$$\bar{R} = \frac{\Sigma R}{k} = \frac{68}{10} = 6.8$$

For $\bar{x}$-chart

$$CL = \bar{\bar{x}} = 46.2$$

$$UCL = \bar{\bar{x}} + A_2\bar{R} = 46.2 + 0.58 \times 6.8$$

$$= 46.2 + 3.944 = 50.144$$

$$LCL = \bar{\bar{x}} - A_2\bar{R} = 46.2 - 0.58 \times 6.8$$

$$= 46.2 - 3.944 = 42.256$$

For R-chart

$$CL = \bar{R} = 6.8$$

$$UCL = D_4\bar{R} = 2.11 \times 6.8 = 14.348$$

$$LCL = D_3\bar{R} = 0 \times 6.8 = 0$$

For $\bar{x}$-chart some of the points are above and below the UCL and LCL, so the process is not under control.

For R-chart all of the points lie within the UCL and LCL, so the process is under control.

Example 4. *A company manufactures screws to a nominal diameter 0.500 ± 0.030 cm. Five samples of size 3 each were taken from the manufactured lot at different lengths. The readings are as follows :*

Sample No.	*Measurement per sample x (in cm.)*		
	1	*2*	*3*
1	*0.488*	*0.489*	*0.505*
2	*0.494*	*0.495*	*0.499*
3	*0.498*	*0.515*	*0.487*
4	*0.492*	*0.509*	*0.514*
5	*0.490*	*0.508*	*0.499*

Calculate the control limits of $\bar{x}$ and R-charts. Comment on the state of control by drawing the charts.

Solution. Calculation for $\bar{\bar{x}}$

NOTES

For sample 1, $\bar{x}_1 = \frac{0.488 + 0.489 + 0.505}{3} = \frac{1.482}{3} = 0.494$

For sample 2, $\bar{x}_2 = \frac{0.494 + 0.495 + 0.499}{3} = \frac{1.488}{3} = 0.496$

For sample 3, $\bar{x}_3 = \frac{0.498 + 0.515 + 0.487}{3} = \frac{1.500}{3} = 0.500$

For sample 4, $\bar{x}_4 = \frac{0.492 + 0.509 + 0.514}{3} = \frac{1.515}{3} = 0.505$

For sample 5, $\bar{x}_5 = \frac{0.490 + 0.508 + 0.499}{3} = \frac{1.497}{3} = 0.499$

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{x}_4 + \bar{x}_5}{5}$$

$$= \frac{0.494 + 0.496 + 0.500 + 0.505 + 0.499}{5} = \frac{2.494}{5} = 0.4988$$

Calculation for $\bar{R}$:

For sample 1, $R_1 = x_{max} - x_{min}$

$R_1 = 0.505 - 0.488 = 0.017$

For sample 2, $R_2 = 0.499 - 0.494 = 0.005$

For sample 3, $R_3 = 0.515 - 0.487 = 0.028$

For sample 4, $R_4 = 0.514 - 0.492 = 0.022$

For sample 5, $R_5 = 0.508 - 0.490 = 0.018$

$$\bar{R} = \frac{R_1 + R_2 + R_3 + R_4 + R_5}{5}$$

$$= \frac{0.017 + 0.005 + 0.028 + 0.022 + 0.018}{5} = \frac{0.090}{5} = 0.018$$

Control limits for $\bar{x}$-chart

$$CL = \bar{\bar{x}} = 0.4988$$

$$UCL = \bar{\bar{x}} + A_2\bar{R} \qquad (\because \quad A_2 = 1.02 \text{ for } n = 3)$$

$$= 0.4988 + 1.02 \times 0.018$$

$$= 0.4988 + 0.01836 = 0.5172$$

$$LCL = \bar{\bar{x}} - A_2\bar{R} = 0.4988 - 1.02 \times 0.018$$

$$= 0.4988 - 0.01836 = 0.4804$$

Control limits for R-chart

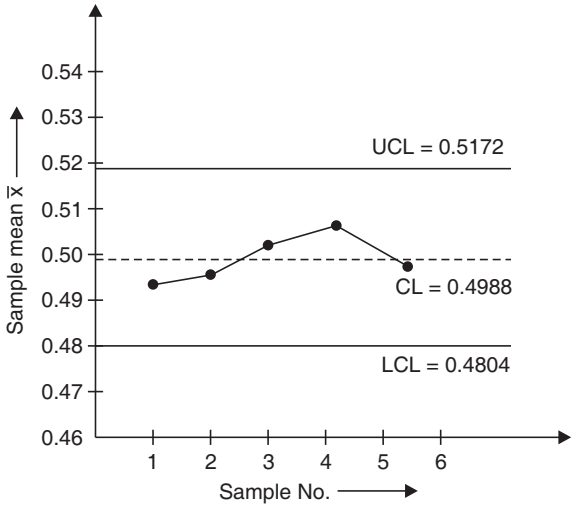
$$CL = \bar{R} = 0.018$$

$$UCL = D_4\bar{R} = 2.57 \times 0.018 \qquad (\because \quad D_4 = 2.57 \text{ for } n = 3)$$

$$= 0.0463$$

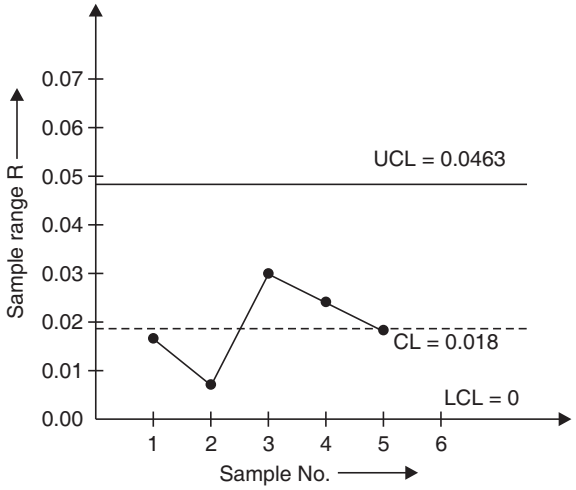
$$LCL = D_3 \overline{R} = 0 \times 0.018 \quad (\because \quad D_3 = 0 \quad \text{for} \quad n = 3)$$
$$= 0$$

NOTES



($\overline{x}$-chart)

It is clear from the figure that all the values of $\overline{x}$ lies within the UCL and LCL, so the process is under control.



(R-chart)

It is clear from the figure that all the values of R lies within the UCL and LCL, so the process is under control.

Example 5. *The following are the mean lengths and ranges of lengths of a finished product from 10 samples each of size 5. The specification limits for length are 200 ± 5 cm. Construct* $\overline{x}$ *and R-charts and examine whether the process is under control.*

Sample No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Mean ($\overline{x}$)	*201*	*198*	*202*	*200*	*203*	*204*	*199*	*196*	*199*	*201*
Range (R)	*5*	*0*	*7*	*3*	*4*	*7*	*2*	*8*	*5*	*6*

Assume for n = 5, $A_2 = 0.577$, $D_3 = 0$ *and* $D_4 = 2.115$.

NOTES

Solution. The specification limits for length are given to be 200 ± 5 cm. Hence, mean μ is known as 200.

$$\bar{R} = \frac{5+0+7+3+4+7+2+8+5+6}{10} = \frac{47}{10} = 4.7$$

Control limits for $\bar{x}$-chart

$$CL = \mu = 200$$

$$UCL = \mu + A_2\bar{R} = 200 + 0.577 \times 4.7$$
$$= 200 + 2.712 = 202.712$$

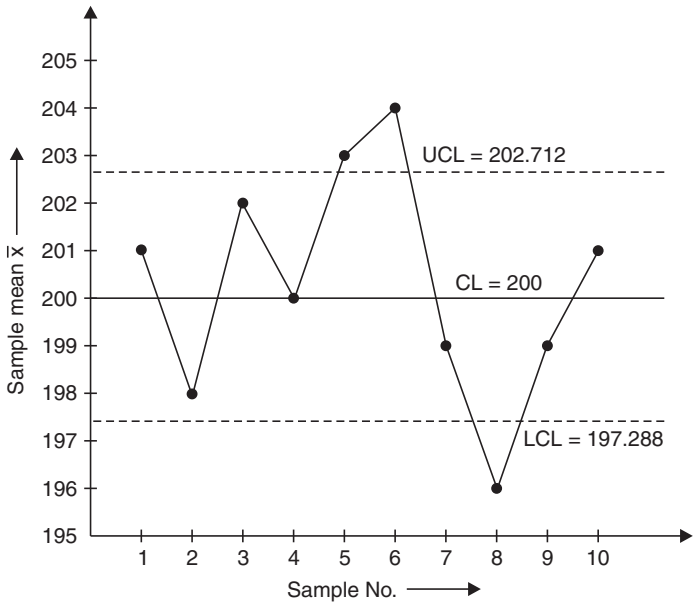
$$LCL = \mu - A_2\bar{R} = 200 - 0.577 \times 4.7$$
$$= 200 - 2.712 = 197.288$$

Control limits for R-chart

$$CL = \bar{R} = 4.7$$

$$UCL = D_4\bar{R} = 2.115 \times 4.7 = 9.941$$

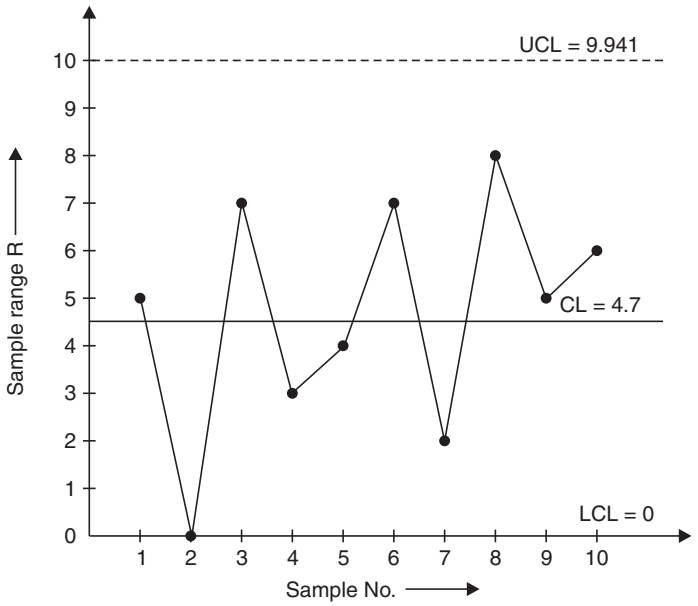
$$LCL = D_3\bar{R} = 0 \times 4.7 = 0$$



($\bar{x}$-chart)

It is clear from figure that three points lie outside the UCL and LCL, so the process is not in control.

NOTES



(R-chart)

It is clear from figure that all the values of R lies within UCL and LCL, so the process is under control.

Example 6. *A bulb manufacturing company ABC samples the fused bulb, taking sample of 5 each every hour. These samples sets of five have been arranged in increasing orders as follows :*

45	*42*	*20*	*35*	*43*	*52*	*61*	*20*	*16*	*70*	*65*	*60*
68	*46*	*25*	*55*	*52*	*70*	*65*	*25*	*28*	*100*	*85*	*75*
75	*65*	*82*	*69*	*57*	*75*	*70*	*32*	*40*	*110*	*95*	*94*
77	*70*	*87*	*78*	*60*	*80*	*90*	*55*	*65*	*115*	*100*	*109*
88	*92*	*87*	*85*	*79*	*120*	*110*	*65*	*85*	*160*	*110*	*140*

Construct $\bar{x}$ *and R-charts and examine whether the process is under control.*

Solution.

For sample 1, $\bar{x}_1 = \dfrac{45 + 68 + 75 + 77 + 88}{5} = \dfrac{353}{5} = 70.6$

Similarly, $\bar{x}_2 = \dfrac{315}{5} = 63$, $\bar{x}_3 = \dfrac{301}{5} = 60.2$, $\bar{x}_4 = \dfrac{322}{5} = 64.4$,

$\bar{x}_5 = \dfrac{291}{5} = 58.2$, $\bar{x}_6 = \dfrac{397}{5} = 79.4$, $\bar{x}_7 = \dfrac{396}{5} = 79.2$

$\bar{x}_8 = \dfrac{197}{5} = 39.4$, $\bar{x}_9 = \dfrac{234}{5} = 46.8$, $\bar{x}_{10} = \dfrac{555}{5} = 111$

$\bar{x}_{11} = \dfrac{455}{5} = 91$, $\bar{x}_{12} = \dfrac{478}{5} = 95.6$

For sample 1, $R_1 = x_{max} - x_{min} = 88 - 45 = 43$

NOTES

Similarly, $R_2 = 50$, $R_3 = 67$, $R_4 = 50$, $R_5 = 36$
$R_6 = 68$, $R_7 = 49$, $R_8 = 45$ $R_9 = 69$
$R_{10} = 90$, $R_{11} = 45$, $R_{12} = 80$

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_{12}}{12} = \frac{858.8}{12} = 71.57$$

$$\bar{R} = \frac{R_1 + R_2 + \ldots + R_{12}}{12} = \frac{692}{12} = 57.67$$

Control limits for $\bar{x}$-chart

$$CL = \bar{\bar{x}} = 71.57$$

$$UCL = \bar{\bar{x}} + A_2\bar{R}$$
$$= 71.57 + 0.577 \times 57.67 \quad (\because \ A_2 = 0.577 \text{ for } n = 5)$$
$$= 71.57 + 33.27 = 104.84$$

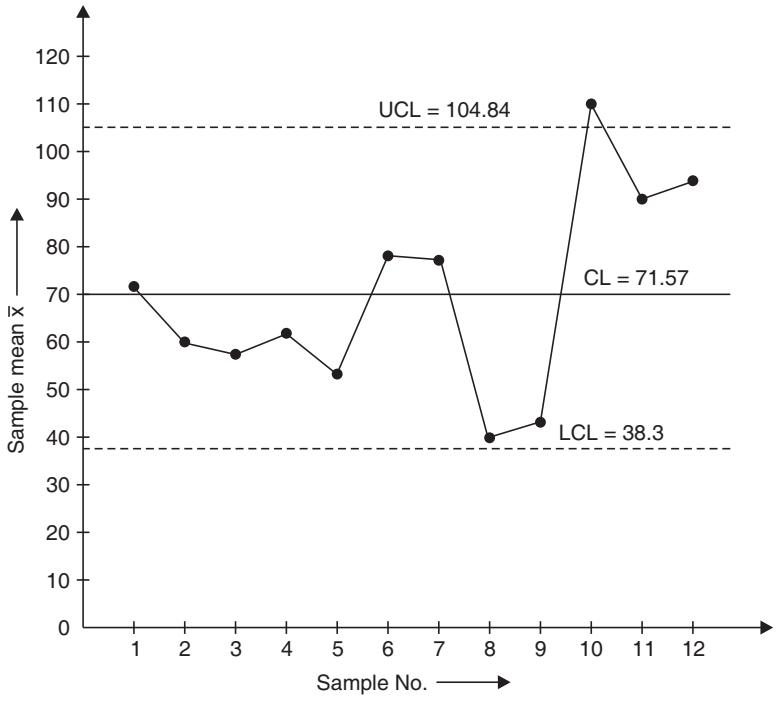
$$LCL = \bar{\bar{x}} - A_2\bar{R}$$
$$= 71.57 - 0.577 \times 57.67$$
$$= 71.57 - 33.27 = 38.3$$

Control limits for R-chart

$$CL = \bar{R} = 57.67$$

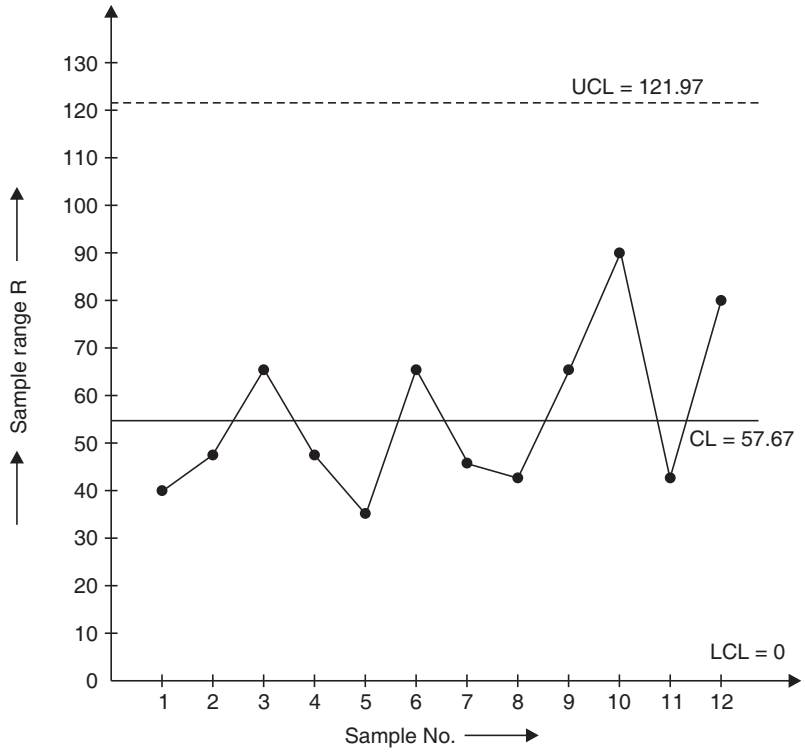
$$UCL = D_4\bar{R}$$
$$= 2.115 \times 57.67 \quad (\because \ D_4 = 2.115 \text{ for } n = 5)$$
$$= 121.97$$

$$LCL = D_3\bar{R}$$
$$= 0 \times 57.67 \quad (\because \ D_3 = 0 \text{ for } n = 5)$$
$$= 0$$



It is clear from the figure that one point lie outside the UCL, so the process is not in control.

NOTES



(R-chart)

It is clear from the figure that all the values of R lies within UCL and LCL, so the process is under control.

Example 7. *In a factory producing spark plugs, the number rejected found in the inspection of 10 lots of size 100 each is given below:*

Lot No.	*Number rejected*	*Fraction rejected*	*Lot No.*	*Number rejected*	*Fraction rejected*
1	*4*	*0.040*	*6*	*4*	*0.040*
2	*7*	*0.070*	*7*	*5*	*0.050*
3	*8*	*0.080*	*8*	*8*	*0.080*
4	*2*	*0.020*	*9*	*6*	*0.060*
5	*3*	*0.030*	*10*	*10*	*0.100*

Construct appropriate control chart and state whether the process is in control.

Solution. Since we are given fraction rejected, p-chart is suitable for the given situation.

$$\bar{p} = \frac{\text{Total number of rejected}}{\text{Total number of items inspected in all samples}}$$

$$= \frac{57}{10 \times 100} = 0.057$$

$$\text{CL}_p = \bar{p} = 0.057$$

NOTES

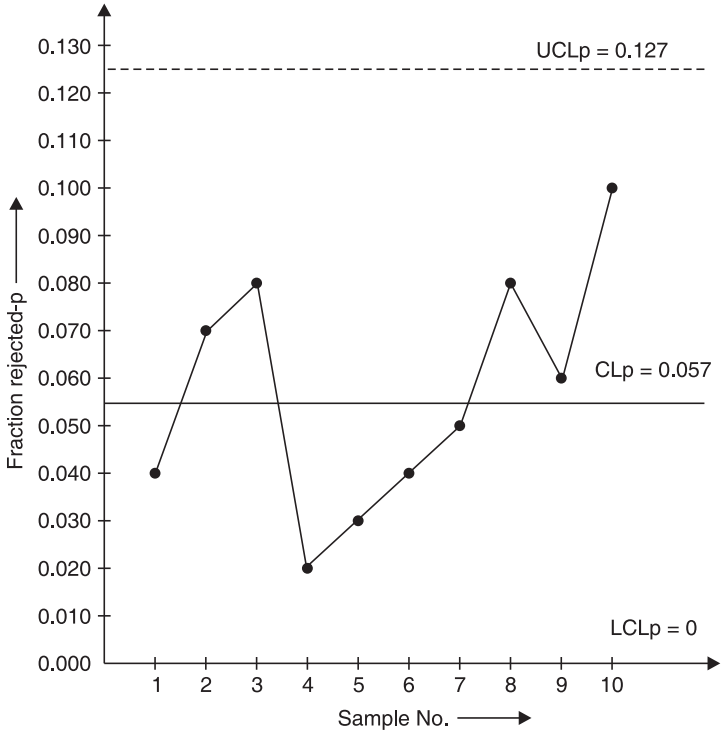
$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.057 + 3\sqrt{\frac{0.057(1-0.057)}{100}}$$

$$= 0.057 + 0.070 = 0.127$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.057 - 3\sqrt{\frac{0.057(1-0.057)}{100}}$$

$$= 0.057 - 0.070 = -0.013$$

Since LCL_p is negative, so LCL_p is taken as zero.



(*p*-chart)

It is clear from figure that all the values of fraction rejected *p* lies within UCL and LCL, so the process is under control.

Example 8. *Based on 15 subgroups each of size 200 taken at intervals of 45 minutes from a manufacturing process, the average fraction defective was found to be 0.068. Calculate the value of CL, UCL and LCL.*

Solution. Since we are given average fraction defective, we will calculate the control limits of *p*-chart.

$$CL_p = \bar{p} = 0.068$$

$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.068 + 3\sqrt{\frac{0.068(1-0.068)}{200}}$$

$$= 0.068 + 0.053 = 0.121$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.068 - 3\sqrt{\frac{0.068(1-0.068)}{200}}$$

$$= 0.068 - 0.053 = 0.015.$$

NOTES

Example 9. *Samples of 100 tubes are drawn randomly from the output of a process that produces several thousand units daily. Sample items are inspected for quality and defective tubes are rejected. The results of 15 samples are as follows :*

Sample No.	*No. of defective tubes*	*Sample No.*	*No. of defective tubes*
1	*8*	*9*	*10*
2	*10*	*10*	*13*
3	*13*	*11*	*18*
4	*9*	*12*	*15*
5	*8*	*13*	*12*
6	*10*	*14*	*14*
7	*14*	*15*	*9*
8	*6*		

Construct a control chart for fraction defective, and examine whether the process is under control.

Solution.

Sample No.	*No. of defective tubes*	*Fraction defective*
1	8	0.08
2	10	0.10
3	13	0.13
4	9	0.09
5	8	0.08
6	10	0.10
7	14	0.14
8	6	0.06
9	10	0.10
10	13	0.13
11	18	0.18
12	15	0.15
13	12	0.12
14	14	0.14
15	9	0.09

$$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total number of items inspected in all samples}}$$

$$= \frac{169}{15 \times 100} = 0.113$$

Control limits for p-chart

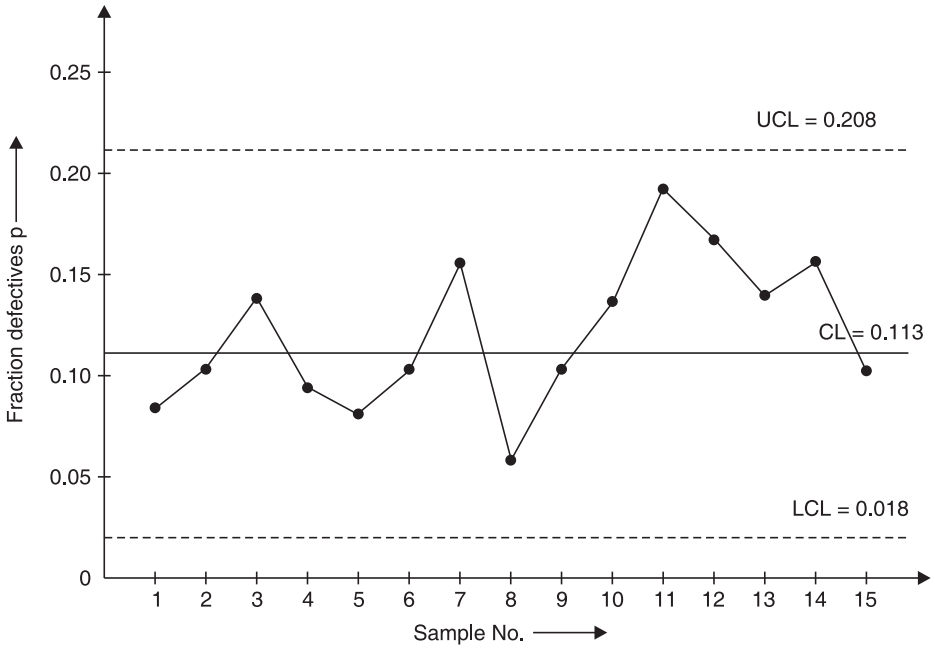
$$CL_p = \bar{p} = 0.113$$

$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.113 + 3\frac{\sqrt{0.113(1-0.113)}}{100}$$

$$= 0.113 + 0.095 = 0.208$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.113 - 3\sqrt{\frac{0.113(1-0.113)}{100}}$$
$$= 0.113 - 0.095 = 0.018$$

NOTES



(*p*-chart)

It is clear from figure that all the values of fraction defectives p lies within UCL and LCL, so the process is under control.

Example 10. *The following data refers to visual defects found during the inspection of the first 10 samples of size 50 each from a lot of two-wheelers manufactured by an automobile company :*

Sample No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
No. of defectives	*4*	*3*	*2*	*3*	*4*	*4*	*4*	*1*	*3*	*2*

Draw the p-chart and examine whether the process is under control.

Solution.

Sample No.	*No. of defectives*	*Fraction defectives*
1	4	0.08
2	3	0.06
3	2	0.04
4	3	0.06
5	4	0.08
6	4	0.08
7	4	0.08
8	1	0.02
9	3	0.06
10	2	0.04

$$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total number of items inspected in all samples}} = \frac{30}{10 \times 50} = 0.06$$

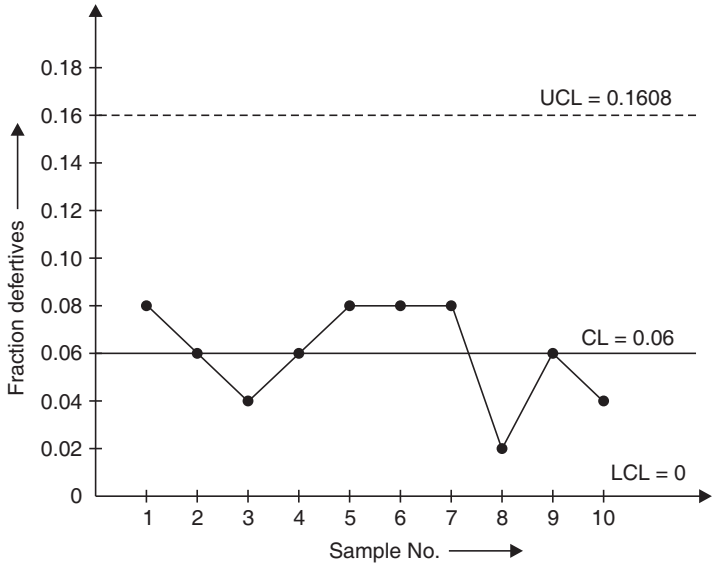
Control limits for p-chart

$$CL_p = \bar{p} = 0.06$$

$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 + 3\sqrt{\frac{0.06(1-0.06)}{50}} = 0.06 + 0.1008 = 0.1608$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 - 3\sqrt{\frac{0.06(1-0.06)}{50}} = 0.06 - 0.1008 = -0.0408$$

Since LCL_p is negative, so $LCL_p = 0$



(p-chart)

It is clear from figure that all the values of fraction defectives p lies within UCL and LCL, so the process is under control.

Example 11. *Ten samples of hourly production of a mass produced items are taken and the number of defectives in each sample are noted. On the basis of these data, obtain the control limits of the control chart for fraction defectives.*

Sample No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Size of sample	*148*	*160*	*155*	*156*	*161*	*167*	*164*	*160*	*156*	*173*
No. of defectives	*7*	*6*	*8*	*8*	*5*	*9*	*8*	*8*	*7*	*10*

Solution. Here, the sample sizes are different. So the average sample size is to be determined first as

Number of items examined = 1600

Number of samples = 10

$$\text{Average sample size } (n) = \frac{1600}{10} = 160$$

NOTES

Now $$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total number of items inspected in all samples}}$$

$$= \frac{76}{1600} = 0.0475$$

Control limits for p-chart

$$CL_p = \bar{p} = 0.0475$$

$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.0475 + 3\sqrt{\frac{0.0475(1-0.0475)}{160}}$$

$$= 0.0475 + 0.0504 = 0.0979$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.0475 - 3\sqrt{\frac{0.0475(1-0.0475)}{160}}$$

$$= 0.0475 - 0.0504 = -0.0029$$

Since LCL_p is negative, so $LCL_p = 0$.

Example 12. *In a blade manufacturing factory, 1000 blades are examined daily. Following information shows number of defectives blades obtained there. Draw the np-chart and comment on the state of control.*

Day	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
No. of defective blades	*9*	*10*	*12*	*8*	*7*	*15*	*10*	*12*	*10*	*8*	*7*	*13*	*14*	*15*	*16*

Solution. Here, $n = 1000$, $k = 15$

$$n\bar{p} = \frac{\text{Total number of defectives}}{\text{Number of samples inspected}}$$

$$n\bar{p} = \frac{166}{15} = 11.067$$

$$\bar{p} = \frac{166}{1000 \times 15} = 0.011$$

Control limits for np-chart

$$CL_{np} = n\bar{p} = 11.067$$

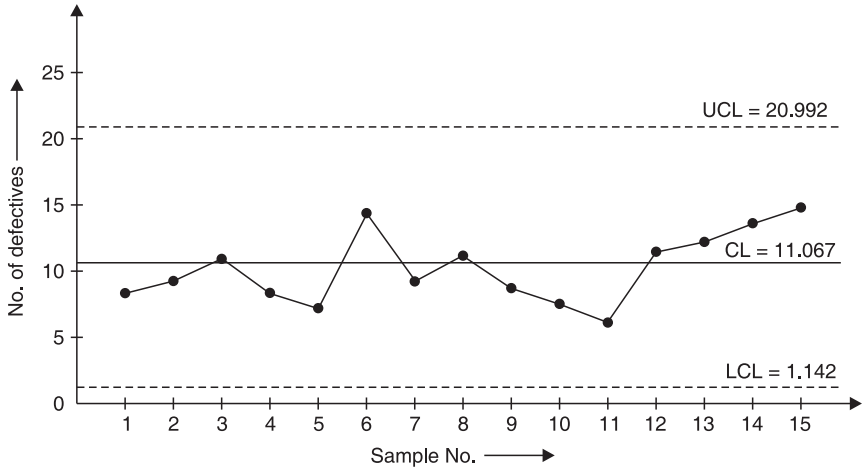
$$UCL_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 11.067 + 3\sqrt{11.067(1-0.011)}$$

$$= 11.067 + 9.925 = 20.992$$

$$LCL_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 11.067 - 3\sqrt{11.067(1-0.011)}$$

$$= 11.067 - 9.925 = 1.142$$

NOTES



(*np*-chart)

It is clear from the figure that all the number of defectives lies within UCL and LCL, the process is under control.

Example 13. *An inspection of 10 samples of size 400 each from 10 lots revealed the following number of defectives 17, 15, 14, 26, 9, 4, 19, 12, 9, 15.*

Draw the control chart for number of defectives and examine whether the process is under control.

Solution. Here, $n = 400$ and $k = 10$

$$n\bar{p} = \frac{\text{Total number of defectives}}{\text{Number of samples inspected}}$$

$$n\bar{p} = \frac{140}{10} = 14$$

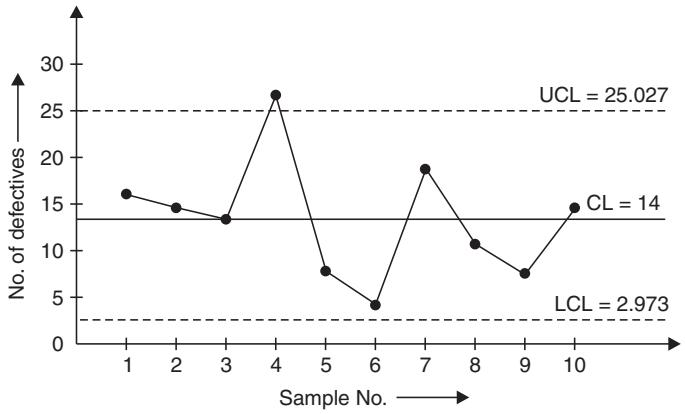
$$\bar{p} = \frac{140}{400 \times 10} = 0.035$$

Control limits for *np*-chart

$$\text{CL}_{np} = 14$$

$$\text{UCL}_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 14 + 3\sqrt{14(1-0.035)}$$
$$= 14 + 11.027 = 25.027$$

$$\text{LCL}_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 14 - 3\sqrt{14(1-0.035)}$$
$$= 14 - 11.027 = 2.973$$



(*np*-chart)

NOTES

It is clear from the figure that one point corresponding to 4th sample lie outside the UCL and LCL, so the process is not under control.

Example 14. *An inspection of 10 samples of size 100 each revealed the following data :*

Sample No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
No. of defectives	*2*	*1*	*1*	*3*	*2*	*3*	*4*	*2*	*2*	*0*

Draw the control chart for number of defectives (np-chart) and examine whether the process is under control.

Solution. Here, $n = 100$, $k = 10$

$$n\bar{p} = \frac{\text{Total number of defectives}}{\text{Number of samples inspected}}$$

$$n\bar{p} = \frac{20}{10} = 2$$

$$\bar{p} = \frac{20}{100 \times 10} = 0.02$$

Control limits for np-chart

$$CL_{np} = n\bar{p} = 2$$

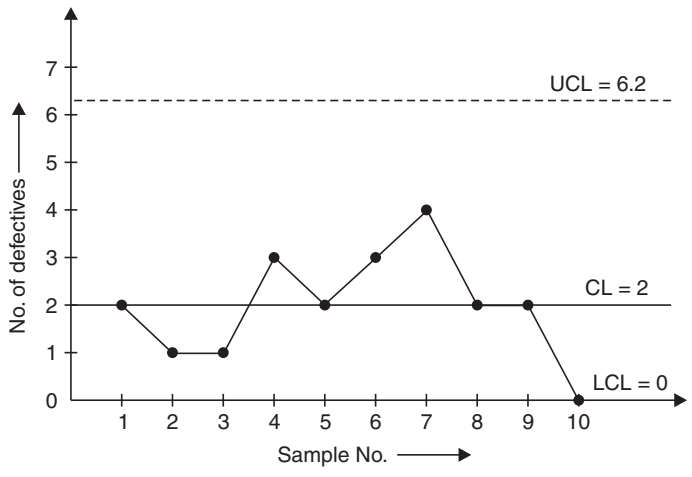
$$UCL_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$= 2 + 3\sqrt{2(1-0.02)} = 2 + 4.2 = 6.2$$

$$LCL_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$= 2 - 3\sqrt{2(1-0.02)} = 2 - 4.2 = -2.2$$

Since LCL_{np} is negative, so $LCL_{np} = 0$.



(np-chart)

It is clear from the figure that all the number of defectives lies within UCL and LCL, the process is under control.

NOTES

Example 15. *During an inspection of equal length of cloth, the following are the number of defects observed :*

2, 3, 4, 0, 5, 6, 7, 4, 3, 2.

Draw a control chart for the number of defects and comment whether the process is under control.

Solution. Average number of defects in 10 sample is given by

$$\bar{c} = \frac{\text{Total number of defects}}{\text{No. of samples inspected}}$$

$$= \frac{2+3+4+0+5+6+7+4+3+2}{10} = \frac{36}{10} = 3.6$$

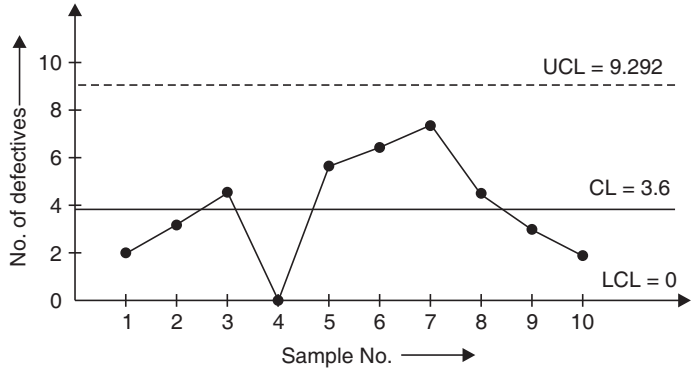
Control limits for c-chart

$$CL_c = \bar{c} = 3.6$$

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}} = 3.6 + 3\sqrt{3.6} = 3.6 + 5.692 = 9.292$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}} = 3.6 - 3\sqrt{3.6} = 3.6 - 5.692 = -2.092$$

Since LCL_c is negative, so $LCL_c = 0$



(c-chart)

It is clear from the figure that all the values of number of defects (c) lies within UCL and LCL, the process is under control.

Example 16. *The number of complaints received daily by an organization are as follows :*

Day	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Complaints	*2*	*3*	*0*	*1*	*9*	*2*	*0*	*0*	*4*	*2*	*0*	*7*	*0*	*2*	*4*

Draw a suitable control chart and examine whether the process is under control.

Solution. For the given problem, the suitable control chart is c-chart. Let the number of complaints is denoted by c.

Here,
$$\bar{c} = \frac{\text{Total number of complaints}}{\text{Number of days}}$$

$$\bar{c} = \frac{36}{15} = 2.4$$

NOTES

Control limits for c-chart

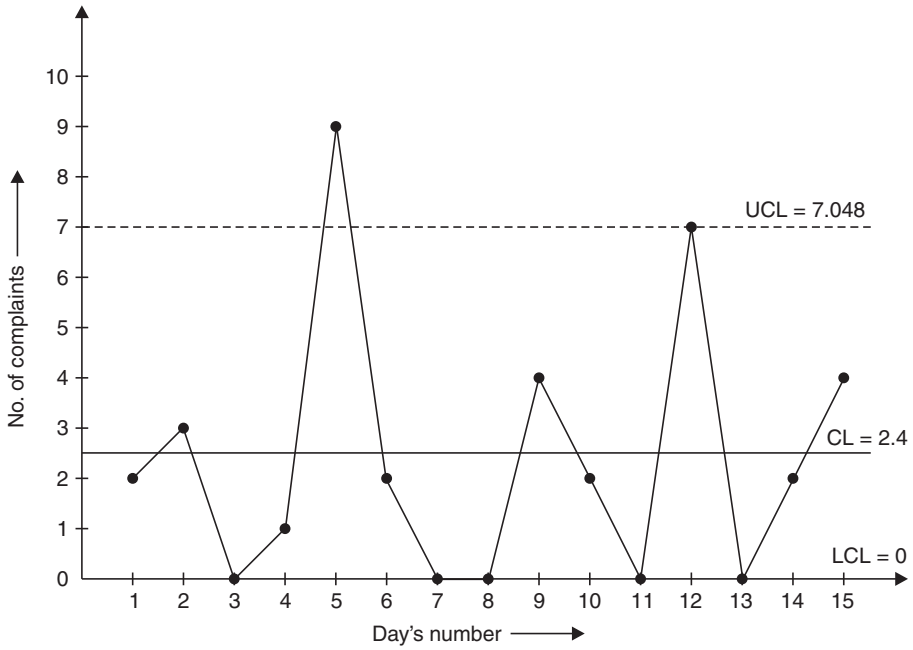
$$CL_c = \bar{c} = 2.4$$

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}} = 2.4 + 3\sqrt{2.4} = 2.4 + 4.648 = 7.048$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}} = 2.4 - 3\sqrt{2.4}$$

$$= 2.4 - 4.648 = -2.248$$

Since LCL_c is negative, so $LCL_c = 0$



It is clear from the figure that one value of c corresponding to 5th day is not within the control limits, the process is out of control.

Example 17. *The following table shows the number of missing rivets observed at the same time of the inspection of 12 aircrafts. Find the control limits for the number of defects chart and comment on the state of control.*

Aircraft Number	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
No. of missing rivets	*7*	*15*	*13*	*18*	*10*	*14*	*13*	*10*	*20*	*11*	*22*	*15*

Solution.

$$\bar{c} = \frac{\text{Total number of missing rivets}}{\text{Number of aircrafts inspected}}$$

$$\bar{c} = \frac{168}{12} = 14$$

Control limits for c-chart

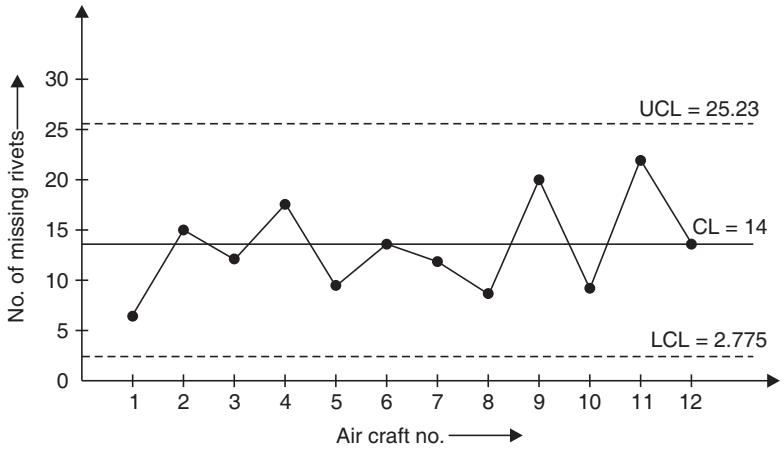
$$CL_c = \bar{c} = 14$$

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}}$$

$$= 14 + 3\sqrt{14} = 14 + 11.225 = 25.23$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}}$$
$$= 14 - 3\sqrt{14} = 14 - 11.225 = 2.775$$

NOTES



(*c*-chart)

It is clear from the figure that all the values of missing rivets lies within UCL and LCL, the process is under control.

EXERCISE 8.1

1. In the manufacturing process of a certain item from 20 subgroups, each of size 4, it is found that $\Sigma\bar{x} = 41.283$ and $\Sigma R = 0.335$. Compute the control limits for $\bar{x}$ and R-charts.

2. A machine is set to deliver packets of a given weight, 10 samples of size 5 each were recorded and the mean and range of each sample is as follows :

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{x}$)	15	17	15	18	17	14	18	15	17	16
Range (R)	7	7	4	9	8	7	12	4	11	5

Calculate the control limits of $\bar{x}$ and $\bar{R}$-charts and comment on the state of control.

3. A company manufactures a product which is packed in cans. It utilises an automatic filling equipment. It takes a sample of 5 cans every hour and measures the filling (grams) in the last 5 samples.

Sample No.	*Individual measurements*				
	1	*2*	*3*	*4*	*5*
1	1001	998	1002	1002	999
2	999	998	1001	998	999
3	995	1001	1003	1002	1002
4	1000	998	999	1001	1002
5	994	1000	996	996	999

Calculate the control limits of $\bar{x}$ and R-charts and comment on the state of control.

NOTES

4. The following data shows the value of sample mean $\bar{x}$ and range R for 10 samples of size 5 each.

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{x}$)	11.2	11.8	10.8	11.6	11.0	9.6	10.4	9.6	10.6	10.0
Range (R)	7	4	8	5	7	4	8	4	7	9

Construct $\bar{x}$ and R-charts and examine whether the process is under control.

(Given for $n = 5$, $A_2 = 0.577$, $D_3 = 0$, $D_4 = 2.115$)

5. The following data shows the value of sample mean $\bar{x}$ and range R for 10 samples of size 5 each.

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{x}$)	43	49	37	44	45	37	51	46	43	47
Range (R)	5	6	5	7	7	4	8	6	4	6

Construct $\bar{x}$ and R-charts and comment on the state of control.

6. If the average fraction defective of large sample of products is 0.1537. Calculate the control limits. (Given that subgroup size is 2000)

7. A company produces fuses for automobile electric systems. Five hundred of the fuses are tested per day for 30 days. The following table gives the number of defective fuses found per day for the 30 days.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of defectives	3	3	3	3	1	1	1	1	6	1	1	1	5	4	6
Day	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No. of defectives	3	6	2	7	3	2	3	6	1	2	3	1	4	4	5

Calculate the central line, upper control limit and lower control limit for *p*-chart.

8. A daily sample of 30 items was taken over a period of 14 days in order to establish attributes control limits. If 21 defectives were found, what should be the upper and lower control limits of the proportion of defectives ?

9. In a factory producing an item, the number rejected found in the inspection of 20 lots of size 100 each is given below :

Lot No.	*No. rejected*	*Fraction rejected*	*Lot No.*	*No. rejected*	*Fraction rejected*
1	5	0.050	11	4	0.040
2	10	0.100	12	7	0.070
3	12	0.120	13	8	0.080
4	8	0.080	14	2	0.020
5	6	0.060	15	3	0.030
6	5	0.050	16	4	0.040
7	6	0.060	17	5	0.050
8	3	0.030	18	8	0.080
9	3	0.030	19	6	0.060
10	5	0.050	20	10	0.100

Construct appropriate control chart and state whether the process is in control.

NOTES

10. The table given below shows the results of the production and inspection of 100 castings a day for 20 days. Based on these data, construct *p*-chart and state whether the process is in control.

Day	*No. of defectives*	*Day*	*No. of defectives*
1	6	11	33
2	11	12	39
3	20	13	25
4	22	14	18
5	9	15	17
6	40	16	14
7	12	17	13
8	10	18	5
9	31	19	7
10	30	20	9

11. The following data refer to defects found during inspection of the first 10 samples of size 100 each.

Sample No.	1	2	3	4	5	6	7	8	9	10
No. of defectives	4	8	11	3	11	7	7	16	12	6

Calculate the control limits for *np*-chart and state whether the process is in control.

12. Twenty samples each of size 10 were inspected. The number of defectives found in each of them is given below :

Sample No.	1	2	3	4	5	6	7	8	9	10
No. of defectives	0	1	0	3	9	2	0	7	0	1
Sample No.	11	12	13	14	15	16	17	18	19	20
No. of defectives	1	0	0	3	1	0	0	2	0	0

Construct appropriate chart and state whether the process is in control.

13. The following data refer to number of defectives found during inspection of first 10 samples of size 100 each :

Sample No.	1	2	3	4	5	6	7	8	9	10
No. of defectives	4	8	11	3	11	7	7	16	12	6

Obtain the upper and lower control limits of *np*-chart and state whether the process is in control.

14. The following data refer to number of defectives found on 24 consecutive production days in daily samples of 400 items :

Production day	1	2	3	4	5	6	7	8	9	10	11	12
No. of defectives	20	10	20	24	22	18	38	8	24	54	50	18
Production day	13	14	15	16	17	18	19	20	21	22	23	24
No. of defectives	24	30	16	28	20	8	22	22	52	6	20	22

Draw *np*-chart and state whether the process is in control.

NOTES

15. Ten pieces of cloth out of different rolls of equal length contained the following number of defects :

1, 7, 3, 1, 2, 4, 8, 2, 0, 3

Draw a control chart for the number of defects and state whether the process is in control.

16. The number of complaints received daily by an organization are as follows :

Day	1	2	3	4	5	6	7	8	9	10
Complaints	2	3	4	0	5	6	7	4	3	2

Draw a control chart for number of defects and comment whether the process is in control.

17. The number of mistakes made by an account clerk are as follows :

Week No.	1	2	3	4	5	6	7	8	9	10
No. of mistakes	1	0	2	0	1	0	1	0	1	2
Week No.	11	12	13	14	15	16	17	18	19	20
No. of mistakes	3	3	1	0	0	7	1	0	1	0

Draw an appropriate control chart and state whether the mistakes of the clerk is in under control.

Answers

1. For $\bar{x}$ -chart For R-chart
CL = 2.06415 CL = 0.01675
UCL = 2.076 UCL = 0.03819
LCL = 2.0522 LCL = 0]

2. For $\bar{x}$ -chart For R-chart
CL = 16.2 CL = 7.4
UCL = 20.492 UCL = 12.3
LCL = 11.908 LCL = 0
Process is under control by both charts.

3. For $\bar{x}$ -chart : CL = 999.32, UCL = 1001.772, LCL = 996.872,
For R-chart : CL = 4.4, UCL = 9.306, LCL = 0
Process under control using both $\bar{x}$ and R-charts.

4. For $\bar{x}$ -chart : CL = 10.66, UCL = 14.295, LCL = 7.025 ; process under control
For R-chart : CL = 6.3, UCL = 13.3245, LCL = 0 ; process under control.

5. For $\bar{x}$ -chart : CL = 44.2, UCL = 47.564, LCL = 40.836; out of control
For R-chart : CL = 5.8, UCL = 12.267, LCL = 0; under control

6. $CL_p = 0.1537$, $UCL_p = 0.17788$, $LCL_p = 0.1295$.

7. CL = 0.184, UCL = 0.236, LCL = 0.132.

8. $CL_p = 0.05$, $UCL_p = 0.17$, $LCL_p = 0$.

9. $CL_p = 0.06$, $UCL_p = 0.1311$, $LCL_p = 0$; out of control.

10. $CL_p = 0.186$, $UCL_p = 0.303$, $LCL_p = 0.069$; out of control.

11. $CL_{np} = 8.5$, $UCL_{np} = 16.87$, $LCL_{np} = 0.13$; under control.

12. $CL_{np} = 1.5$, $UCL_{np} = 4.89$, $LCL_{np} = 0$; out of control.

13. $UCL_{np} = 16.87$, $LCL_{np} = 0.13$; under control.

14. $CL_{np} = 24$, $UCL_{np} = 38.25$, $LCL_{np} = 9.75$; out of control.

15. $CL_c = 3.6$, $UCL_c = 8.38$, $LCL_c = 0$; under control.

16. $CL_c = 3.6$, $UCL_c = 9.292$, $LCL_c = 0$; under control.

17. $CL_c = 1.2$, $UCL_c = 4.49$, $LCL_c = 0$; not under control.