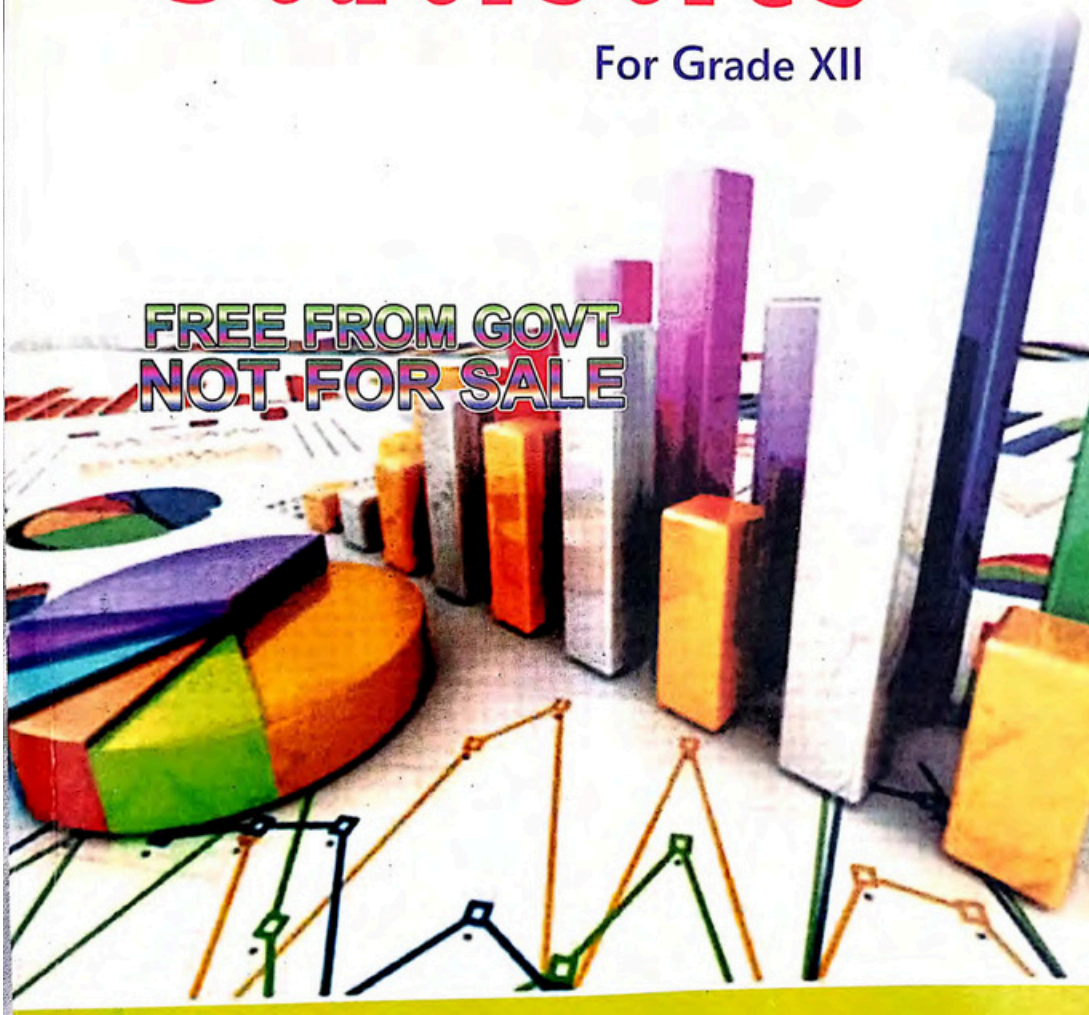
A Textbook of

Statistics

For Grade XII



Khyber Pakhtunkhwa Textbook Board
Peshawar

مومن تو آپس میں بھائی بھائی ہیں۔
پس اپنے دو بھائیوں میں صلح کرادیا کرو
اور اللہ سے ڈرتے رہو تا کہ تم پر رحم
کیا جائے۔
(سورۃ الحجرات:۱۰)

Statistics

for Grade XII

Khyber Pakhtunkhwa Textbook Board
Peshawar

Developed by University Books Printers & Publishers, Peshawar and approved by The Directorate of Curriculum & Teacher Education (DCTE), Abbottabad vide NOC No. 5400-5402/F.03/Vol-I/Statistics-XII dated 08-10-2018.

Author:
Prof. Jamal Shah, Principal, Govt. Degree College, Zaida, Swabi.

Review Supervision:
Gohar Ali Khan
Director, Curriculum & Teacher Education Khyber Pakhtunkhwa
Abbottabad

Focal person:
Zulfiqar Khan
Additional Director, Curriculum & Textbook Review, DCTE
Abbottabad

Reviewers:
- Mr. Zia Shahid, Principal, GHS No. 3, Abbottabad.
- Mr. Majid Khan, Assistant Professor, GPGC No. 1,Abbottabad.
- Mr. Iftikhar Hussain, Senior Subject Specialist GHSS Nawashehr, Abbottabad.
- Ms. Samia Danish, (Desk Officer), DCTE, Abbottabad.
- Mr. Shakeel Iqbal, Subject Specialist, KP Textbook Board, Peshawar

Editor:
Mr. Shakeel Iqbal, Subject Specialist,
Khyber Pakhtunkhwa Textbook Board Peshawar

Printing Supervision:
- Mr. Zakir Hussain Afridi, Chairman
- Mr. Saeedur Rehman, Member (E&P)
  Khyber Pakhtunkhwa Textbook Board, Peshawar

Academic Year 2019-20
Website: www.kptbb.gov.pk
Email: membertbb@yahoo.com
Phone: 091- 9217159-60

Table of Content

Unit - 01

PROBABILITY

After studying this unit, the students will be able to

- Know n! (n factorial) as the notation to express the product of first n natural numbers.
- Describe fundamental principles of counting and illustrate it using tree diagram.
- Explain the meaning of permutation and interpret the number of permutations of n different objects taken r at a time.
- Explain the meaning of combination and interpret the number of combinations of n different objects taken r at a time.
- Define random experiment, sample space, sample point, event, simple and compound events, impossible and sure events, complementary events, equally likely events, exhaustive events, mutually exclusive events.
- Elaborate the term 'probability' through classical definition, relative frequency definition. Axiomatic definition.
- Recognize the formula for probability of occurrence of an event A
- Apply the formula and using Venn diagrams to find the probability in simple cases for the occurrence of an event.
- Describe probability of non-occurrence of an event, odds for the occurrence and odds against the occurrence for an event.
- Recognize the law of probability of complementation.
- State the laws of probability under addition and apply them to solve real life problems.
- Differentiate between dependent and independent events.
- Define the conditional probability and state the laws of probability under multiplication.
- Apply the laws of probability under multiplication to solve real life problems.
- Compute probabilities for real life problems involving counting techniques and probability trees.

1.1 Counting Techniques

Counting rules help us to know about the number of all possible results of an experiment without actually writing them. A few of the commonly used counting rules to solve the probability problems are:

♦ Factorial

The product of first n natural numbers is called factorial and is denoted by the symbol n! (read n factorial). Thus

$1! = 1$

$2! = 2 \cdot 1 = 2$

$3! = 3 \cdot 2 \cdot 1 = 6$

$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$

$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$

$\vdots$

$n! = n\,(n-1)\,(n-2)\ldots 3 \cdot 2 \cdot 1$

This can simply be written as

$n! = n\,(n-1)\,(n-2)!$

$= n\,(n-1)!$

Remember that the quantity $0! = 1$ arbitrarily defined.

♦ Rule of permutation

The groups which can be made from a given number of objects by taking together some or all of the objects at a time with regard to the order of the objects are called permutations. For example, if there are three objects A, B and C, then permutations of size two will be AB, BA, AC, CA, BC, CB. Generally, if we have "n" objects, then total number of permutations of size "r" can be obtained by the formula

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

Consider the above example, n=3 (number of objects), r=2 (group size), therefore the number of permutations is equal to $\overset{3}{P}_{2}=\frac{3!}{(3-2)!}=\frac{3\cdot2\cdot1}{1!}=\frac{3\cdot2\cdot1}{1}=6$

Example 1.1

How many different three digit numbers can be formed from the digits 3, 4, 5, 6, 7, 8.

Solution:

Here $n = 6$, $r = 3$, the number of permutations is given by

$$\overset{6}{P}_{3}=\frac{6!}{(6-3)!}=\frac{6!}{3!}=\frac{6\cdot5\cdot4\cdot3!}{3!}=6\cdot5\cdot4=120$$

♦ Permutations when all objects are taken at a time (without repetition)

When all of the given objects are considered at a time in the formation of groups i.e. $r = n$ then the total number of permutations $= \overset{n}{P}_{n} = n!$

Example 1.2

How many different permutations can be made from the letters of the word "BOXER"?

Solution:

Here $r = n = 5$, therefore, total number of permutations $= 5! = 5\cdot4\cdot3\cdot2\cdot1 = 120$

♦ Permutations when all objects are taken at a time (with repetition)

When all of the given objects are considered at a time in the formation of groups and out of them n_1 are of one kind, n_2 are of second kind… n_k are of the k^{th} kind, then the total number of permutations is given by

$$\binom{n}{n_1,n_2,\cdots,n_k}=\frac{n!}{n_1!n_2!\cdots n_k!}\text{, where } n_1+n_2+\ldots+n_k=n$$

Example 1.3

How many different permutations can be made from the letters of the word "STATISTICS"?

Solution:

Here $n = 10$ (total number of letters), $n_1 = 3$ (number of S's), $n_2 = 3$ (number of T's), $n_3 = 2$ (number of I's), $n_4 = 1$ (number of A's) and $n_5 = 1$ (number of C's). Therefore, the required number of permutations is given by

$$\binom{n}{n_1,n_2,n_3,n_4,n_5}=\frac{n!}{n_1!n_2!n_3!n_4!n_5!}$$

$$\binom{10}{3,3,2,1,1}=\frac{10!}{3!3!2!1!1!}=\frac{10\cdot9\cdot8\cdot7\cdot6\cdot5\cdot3\cdot2\cdot1}{3\cdot2\cdot1\cdot3\cdot2\cdot1\cdot2\cdot1\cdot1\cdot1}=50400$$

♦ Rule of combination

The groups which can be made from a given number of objects by taking together some or all of the objects at a time without regard the order of the objects are called combinations. For example, if there are three objects A, B and C then combinations of size two will be AB, AC, and BC. Generally if there are "n" objects then the total number of combinations of size "r" is given by the formula $\overset{n}{C}_{r}=\frac{n!}{(n-r)!r!}$

By formula the number of combinations for above example is equal to

$$\overset{3}{C}_{2}=\frac{3!}{(3-2)!2!}=\frac{3\cdot2\cdot1}{1!2\cdot1}=3$$

It can be deduced that:

- $\binom{n}{n}=\binom{n}{0}=1$
- $\binom{n}{r}=\binom{n}{n-r}$
- $\binom{n}{1}=\binom{n}{n-1}=n$
- $\binom{n}{r}+\binom{n}{r-1}=\binom{n+1}{r}$

Example 1.4

In how many ways a committee of three men can be selected from seven men?

Solution:

Here $n = 7$, $r = 3$, therefore, the total number of combinations is

$$^{7}C_{3} = \frac{7!}{(7-3)!3!} = \frac{7 \cdot 6 \cdot 5 \cdot 4!}{4!3 \cdot 2 \cdot 1} = 35$$

♦ Rule of addition

If there are two mutually exclusive operations having m and n results respectively, then the two operations combined have $(m + n)$ results.

♦ Rule of multiplication

If an experiment has m results and another independent experiment has n results, then the compound experiment has exactly (mn) results.

1.2 Introduction to probability

In everyday life, we face two situations either we may be sure about the occurrence of an event or not sure (uncertain). If sure, then there is no need of probability but if not sure, then probability is used, which is defined as "the numerical evaluation of occurrence of an event is called probability".

Originally, it was known as "science of gamblers" because its foundation was laid by two French mathematicians Pascal and Fermat in connection with gambling problems but nowadays probability theory has wider applications in almost all areas of learning. It is a base for inferential statistics.

For learning the probability technique, it is necessary to understand the following key terms.

1.2.1 Random experiment

The dictionary meaning of random is "unexpected" or "unpredictable", chosen by chance rather than according to a plan.

An experiment whose results cannot be predicted in advance is called random experiment. For example; toss of a coin, roll of a die, etc.

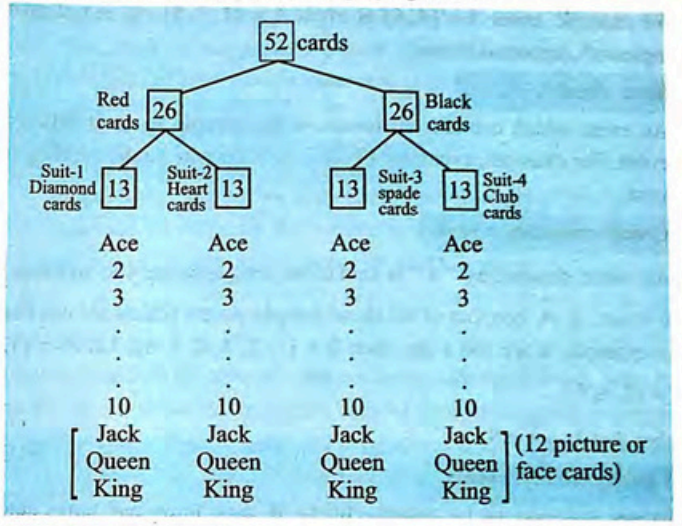
1.2.2 Sample space

The collection of all possible results of a random experiment is called sample space and is denoted by S.

The sample space for tossing a coin is S = {H, T}.

The sample space for rolling a die is S = {1, 2, 3, 4, 5, 6}.

The sample space for an ordinary deck of playing cards is



♦ Definition of outcome

Each result of the sample space is called an outcome or sample point.

1.2.3 Event

The collection of favourable outcomes to a happening from the sample space is called an event or a subset of the sample space is called an event. It is denoted by E_1, E_2 E_3...or A, B, C,...

♦ Impossible event

An event which has no favourable outcome is called an impossible or null event and is denoted by ϕ. For example, A = observing "7" when a die is rolled once. Surely "A" is impossible event because the number 7 cannot happen when an ordinary die is rolled once.

♦ Simple event

An event having exactly one sample point is called simple or elementary event. For example, event A = {5} when a die is rolled, is a simple event.

♦ Compound event

An event which contains more than one outcome is called compound event. For example, event A = {4, 6} or event A = {1, 3, 5} etc. in rolling of a die are examples of compound event.

♦ Sure event

An event which contains all results of the sample space is called sure or certain event. For example, event A= {1, 2, 3, 4, 5, 6} when a die is rolled once, is a sure event.

♦ Complementary event

An event, denoted by "$\overline{A}$" is said to be complementary to an event "A" in a sample space, if $\overline{A}$ consists of all those sample points which are not contained in A. For example, if we roll a die, then S = {1, 2, 3, 4, 5, 6}. Let A = {1, 3, 5}, then $\overline{A}$ = {2, 4, 6}.

Remember that $A \cup \overline{A} = S$.

♦ Equally likely events

Events are said to be equally likely if they have the same chance of occurrence. For example, if we toss a fair coin, then head (H) and tail (T) have same chance of occurrence. So head and tail are equally likely events.

♦ Mutually exclusive events

If two events cannot occur simultaneously then they are called mutually exclusive events. For example, if we toss a coin, H and T cannot occur together so they are mutually exclusive events. Similarly, success and failure, male and female births etc. are mutually exclusive events.

♦ Exhaustive events

Events are called exhaustive if (i) they are mutually exclusive events and (ii) their union makes again the entire sample space.

♦ Dependent events

If the occurrence of an event affects the probability of occurrence of another event, they are called dependent events. Without replacement sampling is an example of dependent events.

♦ Independent events

If the occurrence of an event does not affect the probability of occurrence of any other event, then they are said to be independent events. With replacement sampling is an example of independent events. Similarly, results of two fair coins or examination results of students are independent of each other.

1.2.4 Definitions of probability

♦ The Classical definition of probability

If a sample space has "n" equally likely and mutually exclusive outcomes and if "m" outcomes of them are favourable to the occurrence of an event "A", then the probability of the event A, denoted by $P(A)$, is given by

$$P(A) = \frac{n(A)}{n(S)} = \frac{m}{n}$$

Since events in practical life may not always be equally likely that is why this definition has the shortcoming that it can only be applied to pattern experiments like tossing of a coin, rolling of a die, drawing of a playing card. This definition was given by Laplace. It is also called mathematical or priori definition of probability.

♦ The Relative frequency definition of probability

This definition uses the relative frequency of the occurrence of an event A, over a very large number of trials, that is,

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

This definition, given by Von Mises, has the shortcoming that the experiment needs to be performed a large number of times which is practically time consuming and expensive. It is also called statistical or posteriori definition of probability because it is calculated after conducting the actual experiment.

♦ Axiomatic definition of probability

Axiom means a rule or principle that many people accept as true. To avoid different shortcoming and compute probability, Russian mathematician Kolmogorov imposed some axioms on the probability of an event given below:

Axiom (I) $0 \leq P(E) \leq 1$, where E is any event.

Note that:

(i) If $P(E) = 0$, the event E is said to be null event.

(ii) If $P(E) = 1$, the event E is said to be sure event.

Axiom (II) $P(S) = 1$

Axiom (III) If E_1 and E_2 are mutually exclusive events, then

$P(E_1 \cup E2) = P(E_1) + P(E_2)$

1.2.5 Calculation of probability in case of simple events

Example 1.5

If a coin is tossed, what is the chance of a head?

Solution:

$n(S) = 2^n = 2^1 = 2$

S = {H, T}

Let A: head occurs

A= {H}

n(A) = 1

$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{m}{n} = \frac{1}{2} = 0.5$

It will be more suitable for interpretation to express the answer in percentage.

Example 1.6

A fair coin is tossed twice. What is the probability that exactly one head occurs?

Solution:

$n(S) = 2^n = 2^2 = 4$

S = {HH, HT, TH, TT}

Let A: exactly one head

A= {HT, TH}

$n(A) = 2$

$P(A) = \frac{n(A)}{n(S)} = \frac{2}{4} = \frac{1}{2} = 0.5$

Example 1.7

Three fair coins are tossed once. Find the probability of

(i) exactly two tails (ii) at least two tails (iii) at most 2 heads.

Solution:

$n(S) = 2^n = 2^3 = 8$

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

(i) Let A: exactly two tails

A= {HTT, THT, TTH}

$n(A) = 3$

$P(A) = \frac{n(A)}{n(S)} = \frac{3}{8} = 0.375$

(ii) Let B: at least two tails

B = {HTT, THT, TTH, TTT}

$n(B) = 4$

$P(B) = \frac{n(B)}{n(S)} = \frac{4}{8} = \frac{1}{2} = 0.5$

(iii) Let C: at most 2 heads

C = {HHT, HTH, HTT, THH, THT, TTH, TTT}

$n(C) = 7$

$P(C) = \frac{n(C)}{n(S)} = \frac{7}{8} = 0.875$

Example1.8

If a die is rolled, what is the probability that the number appearing on top is (i) an odd number (ii) an even number less than 5.

Solution:

For die $n(S) = 6^n$

Here $n(S) = 6^1 = 6$

$S = \{1, 2, 3, 4, 5, 6\}$

(i) Let A: an odd number

$A = \{1, 3, 5\}$

$n(A) = 3$

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = 0.5$$

(ii) Let B: an even number less than 5

$B = \{2, 4\}$

$n(B) = 2$

$$P(B) = \frac{n(B)}{n(S)} = \frac{2}{6} = \frac{1}{3}$$

Example 1.9

Two dice are thrown once, what is the probability that (i) the total score is 8 (ii) score is at least 10 (iii) 5 occurs on the second die.

Solution:

For two dice $n(S) = 6^n = 6^2 = 36$

$$S = \begin{Bmatrix} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{Bmatrix}$$

(i) Let A: the total score is 8

$A = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$

$n(A) = 5$

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$$

(ii) Let B: score is at least 10

$B = \{(4, 6), (5, 5), (6, 4), (5, 6), (6, 5), (6, 6)\}$

$n(B) = 6$

$$P(B) = \frac{n(B)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

(iii) Let C: 5 occurs on the second die

$C = \{(1, 5), (2, 5), (3, 5), (4, 5), (5, 5), (6, 5)\}$

$n(C) = 6$

$$P(C) = \frac{n(C)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

Example 1.10

From an ordinary pack of 52 cards, one card is drawn at random. Find the chance of drawing (i) a heart (ii) a red card.

Solution:

$$n(S) = {}^{n}_{r}C = {}^{52}_{1}C = \frac{52!}{(52-1)!1!} = \frac{52 \cdot 51!}{51!1} = 52$$

(i) Let A: card is heart

$n(A) = {}^{13}_{1}C = 13$ (As there are 13 heart cards in an ordinary pack)

$$P(A) = \frac{n(A)}{n(S)} = \frac{13}{52} = \frac{1}{4} = 0.25$$

(ii) Let B: card is red

$n(B) = {}^{26}_{1}C = 26$ (As there are 26 red cards in a pack)

$$P(B) = \frac{n(B)}{n(S)} = \frac{26}{52} = \frac{1}{2} = 0.5 = 50\%$$

Example 1.11

A bag contains 12 identical balls of which 5 white ,4 red and 3 black balls. Two balls are drawn. What is the probability that (i) both are red (ii) one white ball and one black ball?

Solution:

The bag contains (5W+4R+3B) = 12 balls

$$n(S) = \overset{12}{C}_{2} = \frac{12!}{(12-2)!2!}$$

$$= \frac{12 \cdot 11 \cdot 10!}{10!2 \cdot 1} = 66$$

(i) Let A: 2 balls drawn are red

$$n(A) = \overset{4}{C}_{2} = 6,$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{6}{66} = \frac{1}{11}$$

(ii) Let B: one white ball and one black ball

$$n(B) = \overset{5}{C}_{1} \times \overset{3}{C}_{1} = 5 \times 3 = 15$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{15}{66} = \frac{5}{22}$$

Example 1.12

A retailer wishes to buy two mobile sets from a shop having only 10 Samsung and 5 i-Phone mobiles sets. Find the probability that he will buy (i) two Samsung sets (ii) at least one Samsung set (iii) one Samsung and one i-Phone.

Solution:

There are (10 Samsung + 5 i-Phone) = 15 sets

$$n(S) = \overset{15}{C}_{2} = 105$$

(i) Let A: 2 Samsung sets

$$n(A) = \overset{10}{C}_{2} = 45,$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{45}{105} = 0.43$$

(ii) Let B: at least one Samsung set

B = 1Samsung or 2 Samsung sets

= (1 Samsung and 1 i-Phone) or 2 Samsung sets

$$n(B) = \left\{\overset{10}{C}_{1} \times \overset{5}{C}_{1}\right\} + \overset{10}{C}_{2} = (10 \times 5) + 45 = 50 + 45 = 95$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{95}{105} = 0.90$$

(iii) Let C: one Samsung and one i-Phone set

$$n(C) = \overset{10}{C}_{1} \times \overset{5}{C}_{1} = 10 \times 5 = 50$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{50}{105} = 0.48$$

1.3 Complementation law of probability

Statement:

If $\overline{A}$ is the complement of an event A in the sample space then $P(\overline{A}) = 1 - P(A)$

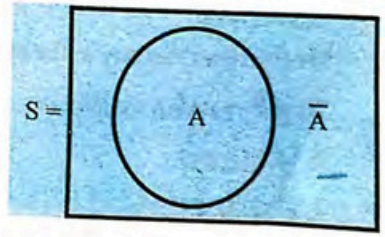
Proof:

As we see from the Venn diagram that event A and $\overline{A}$ are mutually exclusive and exhaustive so $A \cup \overline{A} = S$. Taking probability on both sides

$P(A \cup \overline{A}) = P(S)$

$P(A) + P(\overline{A}) = 1$ (Using axiom II and III)

or $P(\overline{A}) = 1 - P(A)$.



Example 1.13

Five fair coins were tossed once. What is the probability that at least one head occurs?

Solution:

$n(S) = 2^n = 2^5 = 32$

Let A: at least one head.

It will be tedious to list the sample space and pick out the favourable results for at least one head. Alternatively,

$\overline{A}$: No head

$\overline{A}$ = (TTTTT)

$n(\overline{A}) = 1$,

$$P(\overline{A}) = \frac{n(\overline{A})}{n(S)} = \frac{1}{32}$$

Finally, we can obtain the required probability of A by using law of complementation

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{1}{32} = \frac{31}{32} = 0.97$$

Example 1.14

What is the probability that a randomly selected family of four children will have at least one boy?

Solution:

$n(S) = 2^4 = 16$

Let E: at least one boy in the family

Then $\overline{E}$ = no boy i.e. all girls

$\overline{E}$ = {g g g g}

$$P(\overline{E}) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{16}$$

The required probability of E by complementation law is given by

$$P(E) = 1 - P(\overline{E}) = 1 - \frac{1}{16} = \frac{15}{16} = 0.94$$

1.3.1 Addition law of probability for mutually exclusive events

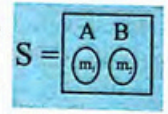
Statement:

If A and B are two mutually exclusive events, then the probability that any one of them happens is equal to the sum of individual probabilities of A and B. In symbol, P(A or B) = P(A ∪ B) = P(A) + P(B)

Proof:

Let n be the total number of sample points in the sample space. Let m_1 be the favourable cases to the occurrence of an event A, then $P(A) = \frac{m_1}{n}$. Let m_2 be the favourable cases to B, so $P(B) = \frac{m_2}{n}$. Since A and B are mutually exclusive events, therefore, favourable cases to (A or B) are equal to m_1+m_2.

Hence $P(A \text{ or } B) = P(A \cup B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B)$



Example 1.15

A card is drawn from an ordinary deck of playing cards. What is the probability that the card will be either a king or a jack?

Solution:

Let the event king be denoted by A and the event jack be denoted by B. These are mutually exclusive events as both cannot occur at a time. Thus we use addition law for mutually exclusive events i.e. $P(A \text{ or } B) = P(A) + P(B)$

Now $n(S) = {}^{52}C_1 = 52$

Let A: card is king

$n(A) = {}^{4}C_1 = 4$

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{52}$$

Let B: card is a jack

$n(B) = {}^{4}C_1 = 4$

$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{52}$$

$$\therefore P(A \text{ or } B) = P(A) + P(B)$$

$$= \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

Example 1.16

A pair of fair dice is rolled once. Find the probability that the sum of the upper dots is either 6 or 9.

Solution:

Let A: Sum of dots is 6 and B: sum of dots is 9. The events are mutually exclusive because they cannot occur together. Hence we use addition rule for mutually exclusive events i.e. $P(A \text{ or } B) = P(A) + P(B)$

Now $n(S) = 6^2 = 36$

$$S = \begin{Bmatrix} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{Bmatrix}$$

$\therefore A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$

$n(A) = 5,$

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$$

$B = \{(3, 6), (4, 5), (5, 4), (6, 3)\}$

$n(B) = 4,$

$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{36}$$

$$\therefore P(A \text{ or } B) = P(A \cup B) = \frac{5}{36} + \frac{4}{36} = \frac{9}{36} = \frac{1}{4}.$$

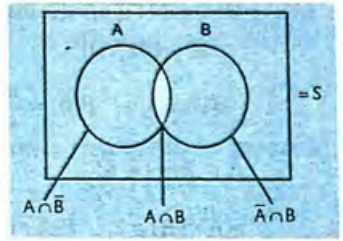
1.3.2 Addition law of probability for not-mutually exclusive events

Statement:

If A and B are two not-mutually exclusive events, then the probability that at least one of the two events A and B occurs is equal to the probability that A occurs plus the probability that B occurs minus the probability that both events A and B occur together. In symbol, $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof:

From the Venn diagram we see that



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

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

$= P(A) + P(\bar{A} \cap B)$, as A and $(\bar{A} \cap B)$ are mutually exclusive.....(i)

Again see the diagram, the set B is given by

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

$P(B) = P[(A \cap B) \cup (\bar{A} \cap B)]$

$= P(A \cap B) + P(\bar{A} \cap B)$, (as $(A \cap B)$ and $(\bar{A} \cap B)$ are mutually exclusive events)

or $P(\bar{A} \cap B) = P(B) - P(A \cap B)$, put in equation (i) we get,

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

Example 1.17

A fair die is thrown once. What is the chance that either an even number or a number greater than 3 will turn up?

Solution:

Let A: an even number occur and B: a number greater than 3 occurs. The events are not mutually exclusive because A may happen, B may happen or both A and B happen. So we use addition law for not-mutually exclusive events:

$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

Now $n(S) = 6^n = 6^1 = 6$

$S = \{1, 2, 3, 4, 5, 6\}$

Let A: an even number occurs

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

$n(A) = 3,\ P(A) = \frac{n(A)}{n(S)} = \frac{3}{6}$

Let B: number is greater than 3

$B = \{4, 5, 6\}$

$n(B) = 3, \quad P(B) = \frac{n(B)}{n(S)} = \frac{3}{6}$

$A \cap B = \{4, 6\}$

$n(A \cap B) = 2, \quad P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{6}$

Hence $P(A \text{ or } B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{3+3-2}{6} = \frac{4}{6} = \frac{2}{3}$

Example 1.18

Ahsan appeared in the annual examination. The probability that he will pass (i) Mathematics is 0.60 (ii) Statistics is 0.50 and (iii) both Mathematics and Statistics is 0.30. What is the probability that Ahsan will pass either Mathematics or Statistics?

Solution:

Let M: Ahsan will pass Mathematics S: Ahsan will pass Statistics

Here the events are not mutually exclusive as Ahsan may pass Mathematics or Statistics or both. So we use addition law for not-mutually exclusive events.

Given that $P(M) = 0.60$

$P(S) = 0.50$

$P(M \cap S) = 0.30$

$\therefore P(M \text{ or } S) = P(M) + P(S) - P(M \cap S)$

$= 0.60 + 0.5 - 0.30 = 0.80$

1.3.3 Multiplication law for independent events

Statement:

If A and B are two independent events, then the probability of their simultaneous happening is equal to the product of their separate probabilities. Symbolically, $P(A \text{ and } B) = P(A \cap B) = P(A)\,P(B)$

Proof:

Let "m" and "n" be the favourable and possible outcomes respectively for an event "A", then $P(A) = \frac{m}{n}$.

Let M and N be the favourable and possible outcomes respectively for an event "B", then $P(B) = \frac{M}{N}$.

Since A and B are independent events, therefore, mM, nN be the favourable and possible outcomes respectively for event (A and B).Thus,

$$P(A \text{ and } B) = P(A \cap B) = \frac{mM}{nN} = \frac{m}{n} \cdot \frac{M}{N} = P(A)\,P(B)$$

Example 1.19

Suppose that a bag contains 10 balls of which 4 are red balls and 6 are green balls. Find the probability of drawing two green balls in succession if the ball that is drawn first is replaced.

Solution:

Let A: first ball drawn is green, B: second ball drawn is green. The events are independent because the ball drawn first is replaced before the next draw, so probability of both green balls will remain the same.

The bag contains (4R + 6G) = 10 balls

$$P(A) = \frac{{}^{6}C_{1}}{{}^{10}C_{1}} = \frac{6}{10}$$

$P(B) = \frac{{}^{6}C_{1}}{{}^{10}C_{1}} = \frac{6}{10}$ .Therefore required probability of both green balls is

$P(A \text{ and } B) = P(A)\,P(B)$

$$= \frac{6}{10} \times \frac{6}{10} = \frac{36}{100} = 0.36$$

1.3.4 Conditional Probability

Many times probability of an event is asked that is conditioned on some available information. For example (i) what is the probability that a person selected at random has diabetes given that he has a family history of diabetes. (ii) What is the probability that 3 occurs on a die given that an odd number has occurred. The given information reduces the original sample space by excluding some outcomes as being impossible which before receiving the information were believed possible. This reduced sample space is called conditional sample space and probabilities associated to it are called conditional probabilities.

♦ Definition of conditional probability

The conditional probability of an event A given that another event B has already occurred is denoted by $P(A/B)$ and is defined as;

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

Example 1.20

A coin is tossed twice. Find the conditional probability that two tails result, given that there is at least one tail?

Solution:

$n(S) = 2^n = 2^2 = 4$

$S = \{HH, HT, TH, TT\}$

Let A: 2 tail appear

$A = \{TT\}$

Let B: at least one tail appears

$B = \{HT, TH, TT\}$

$n(B) = 3$

$$P(B) = \frac{n(B)}{n(S)} = \frac{3}{4}$$

$A \cap B = (TT)$

$n(A \cap B) = 1$

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

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

1.3.5 Multiplication law for dependent events

Statement:

If A and B are two dependent events then the probability that both A and B occur is equal to the probability that A occurs multiplied by the conditional probability of B given that A has already occurred. Symbolically,

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

Proof:

The conditional probability of B given that A has already occurred is

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

Multiplying both sides by $P(A)$,we get

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

or $P(A \cap B) = P(A)\,P(B/A)$ $(\because P(A \cap B) = P(B \cap A)$

Example 1.21

Suppose that a bag contains 10 balls of which 3 are white balls and 7 are green balls. If two balls are drawn at random one after another without replacement, find the probability that both balls drawn are green.

Solution:

Let A: first ball is green, B: second ball is green

Since the first ball drawn is not replaced, therefore, the events are dependent and so we use multiplication law of probability for dependent events i.e.

$P(A \text{ and } B) = P(A)\, P(B/A)$

The bag contains (3W + 7G) = 10 balls

$$P(A) = \frac{{}^{7}C_{1}}{{}^{10}C_{1}} = \frac{7}{10}$$

Since green ball drawn is not replaced, so new position of the bag is (3W+6G) = 9 balls, so

$$P(B/A) = \frac{{}^{6}C_{1}}{{}^{9}C_{1}} = \frac{6}{9}$$

Hence $P(A \text{ and } B) = P(A \cap B) = P(A)\, P(B/A) = \frac{7}{10} \times \frac{6}{9} = \frac{42}{90} = \frac{7}{15}$

Example 1.22

If $P(A) = 0.60$, $P(B) = 0.40$, $P(A \cap B) = 0.24$. (i) What is the relation between A and B? Also find (ii) $P(\overline{A})$ (iii) $P(A \cup B)$ (iv) $P(A/B)$.

Solution:

(i) Here $P(A)\, P(B) = (0.60)\,(0.40) = 0.24 = P(A \cap B)$. Multiplication law for independent events is satisfied. Thus, A and B are independent events.

(ii) By complementation law

$$P(\overline{A}) = 1 - P(A) = 1 - 0.60 = 0.40$$

(iii) $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.60 + 0.40 - 0.24 = 0.76$$

(iii) $$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.24}{0.40} = 0.60 = P(A)$$

(This statement also shows that events A and B are independent).

1.3.6 Odds for the occurrence of an event

The term "odds" refers to the ratio of two probabilities; the ratio of probability of success to the probability of failure. Any kind of probability can be expressed as odds, defined as $d = \frac{p}{1-p}$ where d = odds and p = Probability of success. For example; in throwing a die the probability of 3 is one in six that is $p = \frac{1}{6}$. The odds of getting a 3 are $d = \frac{p}{1-p} = \frac{1/6}{1-1/6} = \frac{1/6}{5/6} = \frac{1}{5}$. That is, odds for 3 are 1 to 5. Odds are used by researchers in different fields.

Example 1.23

A card is drawn from a deck of 52 cards. Find (i) the probability and (ii) the odds that the card drawn will be (a) a diamond? (b) a red card? (c) a king?

Solution:

$$n(S) = {}^{52}C_{1} = 52$$

(a) $$p = \frac{{}^{13}C_{1}}{{}^{52}C_{1}} = \frac{13}{52}$$

$$d = \frac{p}{1-p} = \frac{13/52}{1-13/52} = \frac{13/52}{39/52} = \frac{13}{39}$$

$\Rightarrow 13:39$ or $1:3$

(b) $p = \frac{{}^{26}C_1}{{}^{52}C_1} = \frac{26}{52}$

$$d = \frac{p}{1-p} = \frac{26/52}{1-26/52} = \frac{26/52}{26/52} = 1$$

(c) $p = \frac{{}^{4}C_1}{{}^{52}C_1} = \frac{4}{52}$

$$d = \frac{p}{1-p}$$

$$= \frac{4/52}{1-4/52} = \frac{4/52}{48/52} = \frac{4}{48}$$

$$\Rightarrow 1:12$$

Key points

- Remember that the quantity 0! = 1 arbitrarily defined and 1!=1
- ${}^{n}C_r = \frac{n!}{(n-r)!r!}$
- An experiment whose results cannot be predicted in advance is called random experiment
- The collection of all possible results of a random experiment is called sample space
- The collection of favourable outcomes to a happening from the sample space is called an event.
- If two events cannot occur together, then they are called mutually exclusive events.
- If the occurrence of an event affects the probability of occurrence of another event they are called dependent events.
- If the occurrence of an event does not affect the probability of occurrence of any other event, they are said to be independent events.
- $P(A) = \frac{n(A)}{n(S)} = \frac{m}{n}$
- $0 \leq P(E) \leq 1$, where E is any event.
- If $\overline{A}$ is the complement of an event A in the sample space then $P(\overline{A}) = 1 - P(A)$
- $P(A \text{ or } B) = P(A) + P(B)$ if A and B are mutually exclusive events.
- $P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$, if A and B are not- mutually exclusive events.
- $P(A \text{ and } B) = P(A \cap B) = P(A)\,P(B)$, if A and B are independent
- $P(A/B) = \frac{P(A \cap B)}{P(B)}$, If A and B are dependent.

Exercise

1.1 Write T for true and F for false statement.

i. The range of probability is from zero to one.
ii. An event which contains only one sample point is called compound event.
iii. Events which cannot occur at a time are called not mutually exclusive events.
iv. When a die is rolled four time, the number of sample points in the sample space will be 1296.
v. Two events are mutually exclusive if they have no outcomes in common.
vi. A and B are mutually exclusive events if $P(A \text{ and } B) = P(A)\,P(B)$
vii. When two events are independent, the occurrence of one event will not change the probability of the second event.
viii. The probability of drawing a red card form pack of 52 cards is $\frac{26}{52}$.
ix. Probability of an event will never be negative.
x. The complementary events are always not-mutually exclusive events.

1.2 Fill in the suitable word in the blanks.

(i) An event which cannot occur is known as _____ event.
(ii) Totality of all possible outcomes of a random experiment is called_____
(iii) An orderly arrangements of r distinct things out of n are called _____
(iv) The limits of probability are from _____
(v) The need for probability was originally felt in _____
(vi) $P(A \cup B)$ can be expressed by the _____ law of probability.
(vii) If two events A and B are disjoint, the $P(A \cup B) =$ _____
(viii) If $P(A \cap B) = P(A)\,P(B)$, then events A and B are _____
(ix) If $\bar{A}$ is the compliment of A, the probability of $\bar{A}$ is equal to _____
(x) The probability of obtaining a total of 7 in single throw of two dice is _____

1.3 Choose the correct answer.

(i) The probability of an event A lies between
(a) –1 and +1 (b) –1 and 0
(c) 0 and 1 (d) +1 and –1

(ii) Probability is expressed as
(a) ratio (b) proportion
(c) percentage (d) all of the above

(iii) When two events cannot happen simultaneously in a single trial, the events are said to be
(a) dependent (b) equally likely
(c) mutually Exclusive (d) independent

(iv) An event consisting of those elements, which are not in A is called
(a) primary event (b) derived event
(c) simple event (d) complementary event

(v) If A is an event, the conditional probability of A given A is equal to
(a) zero (b) one
(c) ∞ (d) 0.8

(vi) If a coin is tossed three times, then the probability of getting at most one head is equal to
(a) 3/8 (b) 7/8
(c) 1/2 (d) 1/8

(vii) The probability of throwing an even sum with two fair dice is
(a) 1/4 (b) 1/16
(c) 1 (d) 1/2

(viii) The probability of six on a fair die is 1/6. The probability of not six is
(a) 2/6 (b) 5/6
(c) 1/2 (d) 6/6

(ix) $P(A) + P(\bar{A})$ is equal to
(a) 0 (b) ∞
(c) 1 (d) 0.5

(x) A fair die and a fair coin are thrown at a time. The number of combined outcomes is
(a) 6 (b) 12
(c) 2 (d) 8

1.4 Write short notes on factorials, permutations and combinations.

1.5 Find the values of (i) $\overset{8}{P}_{7}$, (ii) $\overset{25}{P}_{5}$, (iii) $\overset{24}{C}_{4}$, (iv) $\overset{19}{C}_{4}$, (v) $\overset{20}{C}_{11}$

1.6 Find all possible (i) permutations and (ii) combinations of 2 letters chosen from the four litters A, B, C, D.

1.7 How many permutations can be made of the letter of the word "TRIANGLE"? How many of these will begin with T and end with E?

1.8 In how many ways can 5 people be seated on a sofa if there are only three seats available?

1.9 How many permutations can be formed from the letters of the words?
(i) MATHEMATICS (ii) MISSISSIPPIANS (iii) INTERMEDIATE
(iv) EXAMINATION (v) ABBOTTABAD

1.10 Out of 12 books in how many ways can a selection of 5 are made when one specified book is always included.

1.11 Give in brief the concept of probability.

1.12 Explain the following terms.
(i) Random experiment (ii) sample space (iii) outcomes (iv) event (v) impossible event (vi) sure event.

1.13 Differentiate between;
(i) Mutually exclusive events and not mutually exclusive events.
(ii) Independent events and dependent events.
(iii) Probability and conditional probability.

1.14 Define equally likely events, compound events, and exhaustive events.

1.15 Find the probability that an even number appears when a perfect cubical die is rolled.

1.16 When a pair of fair dice is thrown. Find the probability that the sum of 8 appears.

1.17 A fair coin is tossed three times. Find the probability that (i) no head occurs (ii) exactly two head occurs (iii) at least one head occurs.

1.18 From a pack of 52 cards, two cards are drawn randomly. What is the probability that they are king?

1.19 From a deck of 52 cards you are dealt one face drawn what is the chance that the card will turn out to be (i) a club? (ii) a black card? (iii) an ace?

1.20 Of 10 eggs in a refrigerator, 2 are bad. From these 4 eggs are chosen at random. Find the probabilities that (i) all are good (ii) 2 are bad.

1.21 (a) State and prove addition law of probability for not mutually exclusive events.

(b) Abid can solve 60% problems and Ali can solve 80% problems in a book. A problem is chosen at random from this book. What is the probability that Abid or Ali can solve?

1.22 A pair of fair dice is rolled. Find the probability of a sum of either 7 or 11.

1.23 State and prove multiplication laws of probability for independent and dependent events.

1.24 Two coins are tossed at a time. What is the probability of getting a head on the first coin and a tail on second coin?

1.25 From a well-shuffled deck of 52 playing cards, two cards are drawn randomly. What is the probability that both are queens if the first card is (i) replaced, (ii) not replaced.

1.26 (a) What is conditional probability?

(b) If two balanced dice are rolled. Find the conditional probability that sum of dots will be 7, given that it is odd.

1.27 Suppose that A and B are independent events, with $P(A) = 0.6$ and $P(B) = 0.2$ Find (i) $P(A \text{ and } B)$ (ii) $P(A/B)$ (iii) $P(A \text{ or } B)$.

1.28 A problem of statistics is given to 3 students A, B and C whose chances of solving it are $1/2$, $1/3$ and $1/4$ respectively. What is the probability that the problem will be solved?

1.29 In an interview 3 persons A, B and C attended. The chances of being selected for that post is, for A $=1/6$, for B $= 1/5$ and for C $=1/7$.
(i) What is the probability of being selected all the three persons A, B and C.
(ii) What is the probability of being not selected A, B and C respectively.

1.30 Two coins are tossed. What are the probability and the odds that (a) exactly one head occur? (b) Tail occurs on both coins?

Unit - 02

Random Variables and Probability Distributions

After studying this unit, the students will be able to

- Define random variable and differentiate between discrete and continuous random variables with real life examples.
- Describe probability distribution of a discrete random variable
- Find probability distribution of a discrete random variable
- Recognize probability mass function and its properties
- Describe and find the probability distribution of a function of discrete random variable.
- Define and find the expected value of a discrete random variable.
- Find the expected value of a linear function of a discrete random variable.
- Describe and verify properties of expected value of a discrete random variable.
- Apply the properties of expected value of a discrete random variable.
- Define variance and standard deviation of a discrete random variable
- Find mean, variance and standard deviation of a discrete random variable.
- Define and find variance and standard deviation of a linear function of a discrete random variable.
- Describe, verify and apply the properties of variance and standard deviation of a discrete random variable.
- Define probability distribution and probability density function of a continuous random variable.
- Define and expected value, variance and standard deviation of a continuous random variable.
- Describe the properties about the expected value and variance for the sum/difference of two independent random variable x and y also apply these properties to solve real life problems.

2.1 Random variable

Probability distributions, which shall be studied in the coming units, are very easy methods for calculation of probabilities in respective situations but they need quantification of the outcomes of a random experiment. For this purpose, sample space of a random experiment is expressed in numerical form according to a characteristic of interest. This numerical presentation of the sample space is termed as random variable (r.v). It is denoted by English letters X, Y, Z or X_1, X_2, X_3...

2.1.1 Definition of random variable

A random variable is a numerical description of a random experiment.

2.1.2 Types of random variable

Random variable can be classified as follow:

(i) Discrete random variable

A variable which takes jumping values or isolated values is called discrete random variable. For example, number of rotten tomatoes in a crate, number of children per house in a street etc. Its probability distribution is called discrete probability distribution.

(ii) Continuous random variable

A variable which takes any value between two limits, $[a, b]$, $a<b$, is called continuous random variable. For example, life of a mobile set, speed of a car etc. It is written as $a \leq X \leq b$. Its probability distribution is called continuous probability distribution.

2.2 Probability distribution of a discrete random variable

If all possible values of a random variable along with their respective probabilities are shown in tabular form and sum of probabilities is equal to one, then it is called probability distribution. For example, discrete probability distribution of X is presented as

Discrete probability

X	$p(x)$
x_1	$p(x_1)$
x_2	$p(x_2)$
.	.
.	.
.	.
x_n	$p(x_n)$
Total	$\sum_{i=1}^{n} p(x_i) = 1$

Example 2.1

Three children were born in a government hospital on Sunday. Find the probability distribution for the number of girls.

Solution:

Here $n(S) = 2^n = 2^3 = 8$

S = {BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG}

X: the number of girls

= {0, 1, 1, 2, 1, 2, 2, 3}

Probability distribution for X

X	$p(x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$
Total	1

Example 2.2

A fair die is rolled once. Find the probability distribution for up turned faces.

Solution:

When a die is rolled we may get 1 or 2 or 3 or 4 or 5 or 6 each with same probability 1/6 of its occurrence because the die is fair. It can be shown in tabular form, called probability distribution as follows:

X	1	2	3	4	5	6	Total
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

2.2.1 Probability mass function

If probability distribution of a random variable X is expressed in a mathematical form or formula, then it is called probability mass function or probability function. It is denoted by $p(x_i)$ and is presented as:

$$p(x_i) = \begin{cases} p(X = x_i), & i = 1, 2, ..., n \\ 0 & otherwise \end{cases}$$

Remember that the objective of probability distribution and probability function is same. Some writers make no distinction and they use probability distribution and probability function interchangeably.

2.2.2 Properties of discrete probability distribution

A probability distribution and probability mass function $p(x_i)$ must satisfy the following two properties:

(i) $0 \leq p(x_i) \leq 1$

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

First property means that answer of probability must be within the range 0 to 1 (inclusive) and second property means that the sum of probabilities for all possible values of a random variable must be equal to one.

Example 2.3

A discrete random variable has the probability function

$$p(x)=\begin{cases} {}^{3}C_{x}\left(\frac{1}{2}\right)^{3} & for \quad x=0,1,2,3 \\ 0 & otherwise \end{cases}$$

i) Compute probabilities for all values of X.
ii) Check that this is a probability mass function.
iii) Find the probability distribution of X.

Solution:

i) Given $p(x)={}^{3}C_{x}\left(\frac{1}{2}\right)^{3}, x=0,1,2,3$

Putting values of x, we get

$$p(0)={}^{3}C_{0}\left(\frac{1}{2}\right)^{3}=(1)\left(\frac{1}{8}\right)=\frac{1}{8}$$

$$p(1)={}^{3}C_{1}\left(\frac{1}{2}\right)^{3}=(3)\left(\frac{1}{8}\right)=\frac{3}{8}$$

$$p(2)={}^{3}C_{2}\left(\frac{1}{2}\right)^{3}=(3)\left(\frac{1}{8}\right)=\frac{3}{8}$$

$$p(3)={}^{3}C_{3}\left(\frac{1}{2}\right)^{3}=(1)\left(\frac{1}{8}\right)=\frac{1}{8}$$

ii) All probabilities are lying between 0 and 1, and

$$\sum_{x=0}^{3} p(x_i)=p(0)+p(1)+p(2)+p(3)=\frac{1}{8}+\frac{3}{8}+\frac{3}{8}+\frac{1}{8}=1$$

Since both properties are satisfied, therefore, the given formula is a probability mass function or probability function.

iii) The probability distribution of X

X	0	1	2	3	Total
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Example 2.4

Suppose a discrete probability distribution of random variable X is given in the following table:

X	-1	0	1
$p(x)$	$3c$	$3c$	$6c$

Find (i) the value of c (ii) $P(X=0)$ (iii) $P(X<0)$ (iv) $P(X\geq -1)$

Solution:

(i) As $\sum_{x=-1}^{1} p(x)=1$

or $p(-1)+p(0)+p(1)=1$

$$3c+3c+6c=1$$

$$12c=1 \quad \Rightarrow c=\frac{1}{12}$$

Putting the value of c in the given probability distribution, we get

X	-1	0	1	Total
$p(x)$	$\frac{3}{12}$	$\frac{3}{12}$	$\frac{6}{12}$	1

(ii) We see that $P(X=0)=\frac{3}{12}$

(iii) $P(X<0)=P(X=-1)=\frac{3}{12}$

(iv) $P(X\geq -1)=P(X=-1)+P(X=0)+P(X=1)$

$$=\frac{3}{12}+\frac{3}{12}+\frac{6}{12}=\frac{12}{12}=1$$

Example 2.5

Find the value of k so that the function can serve as a probability function of the random variable Y

$p(y)=\begin{cases} ky, & y=1,2,3,4,5 \\ 0, & otherwise \end{cases}$. Also find (i) $P(Y=5)$ (ii) $P(Y>3)$

Solution:

As we know that

$$\sum_{y=1}^{5} p(y)=1$$

$$\sum_{y=1}^{5} ky=1$$

$$k\sum_{y=1}^{5} y=1$$

$$k[1+2+3+4+5]=1$$

$$k(15)=1 \quad \Rightarrow \quad k=\frac{1}{15}$$

Put value of k in the given function we get

$$p(y)=\begin{cases} \frac{1}{15}y, & y=1,2,3,4,5 \\ 0, & otherwise \end{cases}$$

(i) Putting $Y=5$ in the formula, we have $P(Y=5)=\frac{5}{15}$

(ii) $P(Y>3)=P(Y=4)+P(Y=5)=\frac{4}{15}+\frac{5}{15}=\frac{9}{15}$

Example 2.6

A pair of fair dice is rolled. Find (i) the probability distribution for the sum of dots. (ii) Using the probability distribution, compute the probabilities of (a) sum of dots is equal to 7 (b) sum of dots is less than 6 (c) sum of dots is greater than or equal to 2 but less than 5 (d) sum of dots is 5 or 10.

Solution:

(i) $n(S)=6^n=6^2=36$

The sample space for the experiment is:

$$S=\begin{Bmatrix} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{Bmatrix}$$

Let X is a random variable denoting the sum of dots on the upper faces of the two dice. Its probability distribution is:

X	2	3	4	5	6	7	8	9	10	11	12	Total
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

This distribution can also be shown by the formula called probability function as

$$p(x)=\begin{cases}\frac{6-|7-x|}{36}, & x=2,3,4,\ldots 12\\ 0 & \textit{otherwise}\end{cases}$$

(ii) Now using the probability distribution, we get:

a) $P(X=7)=\frac{6}{36}$

b) $P(X<6)=P(X=5)+P(X=4)+P(X=3)+P(X=2)$

$$=\frac{4}{36}+\frac{3}{36}+\frac{2}{36}+\frac{1}{36}=\frac{10}{36}.$$

c) $P(2\le X<5)=P(X=2)+P(X=3)+P(X=4)$

$$=\frac{1}{36}+\frac{2}{36}+\frac{3}{36}=\frac{6}{36}$$

d) $P(X=5 \text{ or } 10)=P(X=5)+P(X=10)$,

$$=\frac{4}{36}+\frac{3}{36}=\frac{7}{36}$$

2.2.3 Probability distribution of a function of discrete random variable

If X is a random variable, then its functions like $X^2, \frac{1}{X}$, $aX + b$ etc. are also random variables and thus have probability distributions. In case of two random variables, say X and Y, their functions like $X + Y$, $X - Y$, $aX + bY$ etc. are also random variables where a and b are any two non-zero constants. The function of a random variable is usually denoted by $H(X)$.

Example 2.7

If a discrete random variable X has the following probability distribution.

X	-2	2	1
$p(x)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Find probability distribution for X^2, $2X + 4$.

Solution:

The probability distribution of the random variable X^2 is

$H(X)=X^2$	$p(x)$
4	$\frac{1}{3}$
4	$\frac{1}{2}$
1	$\frac{1}{6}$
Total	1

The probability distribution of the random variable $2X + 4$ is

H(X) =2X+4	$p(x)$
0	$\frac{1}{3}$
8	$\frac{1}{2}$
6	$\frac{1}{6}$
Total	1

2.2.4 Mathematical expectation of a random variable

Hope you have understood random variable and its presentation methods like probability distribution and probability function. Now we want to study its properties like mean, variance and standard deviation etc.

2.2.5 Definition of mathematical expectation or expected value of a random variable

Mathematical expectation or expected value of a random variable is defined as "the mean of a random variable over a very large number of trails". If X is a discrete random variable having the following probability distribution

X	$p(x)$	$X\,p(x)$
x_1	$p(x_1)$	$x_1\,p(x_1)$
x_2	$p(x_2)$	$x_2\,p(x_2)$
:	.	.
.	.	.
.	.	.
x_n	$p(x_n)$	$x_n\,p(x_n)$
Total	1	$\sum_{i=1}^{n} x_i p(x_i)$

The expected value of X is denoted by E(X) and its formula is given by $E(X)=\sum_{i=1}^{n} x_i p(x_i)$, provided it exists.

Example 2.8

A random variable X has the following probability distribution

X	0	1	2	3
$p(x)$	1/7	3/7	2/7	1/7

Find the mean or expected value of X, i.e. $E(X)$.

Solution:

X	$p(x)$	$X\,p(x)$
0	1/7	0
1	3/7	3/7
2	2/7	4/7
3	1/7	3/7
Total	1	10/7

$$\text{Mean} = E(X) = \sum_{x=0}^{3} x_i\, p(x_i) = \frac{10}{7} = 1.43$$

Example 2.9

A discrete random variable can have the values $x_1 = 3$, $x_2 = 8$, and $x_3 = 10$ and the respective probabilities are 0.2, 0.7 and 0.1. Determine the mean.

Solution:

By definition $E(X) = \sum_{i=1}^{n} x_i p(x_i)$

$$= x_1p(x_1) + x_2p(x_2) + x_3p(x_3)$$

$$= 3(0.2) + 8(0.7) + 10(0.1)$$

$$E(X) = 0.6 + 5.6 + 1 = 7.2$$

2.2.6 Mathematical expectation of a function of discrete random variable

If function of a random variable is denoted by $H(X)$, then expected value of the function is denoted by $E[H(X)]$ and its formula is given by $E[H(X)] = \Sigma H(x_i)p(x_i)$.

Example 2.10

A discrete random variable X has the probability distribution:

X	-2	1	3
$p(x)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

Find (i) $E(X^2)$ (ii) $E(2X+5)$

Solution:

Since X^2 and $(2X+5)$ are functions of the given random variable X, therefore, we first find their probability distributions and then the means as follows:

X	$p(x)$	$H(X)=X^2$	$H(X)=2X+5$	$X^2p(x)$	$(2X+5)p(x)$
-2	$\frac{1}{3}$	4	1	$\frac{4}{3}=1.33$	$\frac{1}{3}=0.33$
1	$\frac{1}{6}$	1	7	$\frac{1}{6}=0.17$	$\frac{7}{6}=1.17$
3	$\frac{1}{2}$	9	11	$\frac{9}{2}=4.5$	$\frac{11}{2}=5.5$
Total	1	-	-	6	7

Now $E\{H(X)\} = \Sigma H(x_i)\, p(x_i)$

(i) $E\{X^2\} = \Sigma x_i^2\, p(x_i) = 6$

(ii) $E\{2X+5\} = \Sigma(2x_i+5)\, p(x_i) = 7$

Example 2.11

The probability function of a discrete random variable Y is given by

$$p(y) = \begin{cases} \binom{3}{y}\left(\frac{1}{2}\right)^y\left(\frac{1}{2}\right)^{3-y}, & y = 0, 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}$$

Find $E(Y)$ and $E(Y^2)$.

Solution:

By definition

$$E(Y) = \sum_{y=0}^{3} y_i p(y_i)$$

$$= \sum_{y=0}^{3} y \binom{3}{y}\left(\frac{1}{2}\right)^y\left(\frac{1}{2}\right)^{3-y}$$

$$= 0\binom{3}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{3-0} + 1\binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{3-1} + 2\binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{3-2} + 3\binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{3-3}$$

$$= 0 + 3\left(\frac{1}{8}\right) + 6\left(\frac{1}{8}\right) + 3\left(\frac{1}{8}\right) = \frac{12}{8} = 1.5$$

$$E(Y^2) = \sum_{y=0}^{3} y^2 \binom{3}{y}\left(\frac{1}{2}\right)^y\left(\frac{1}{2}\right)^{3-y}$$

$$= 0\binom{3}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{3-0} + 1\binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{3-1} + 4\binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{3-2} + 9\binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{3-3}$$

$$= 0 + 3\left(\frac{1}{8}\right) + 12\left(\frac{1}{8}\right) + 9\left(\frac{1}{8}\right) = \frac{24}{8} = 3$$

2.2.7 Properties of mathematical expectation

Expected value of a random variable satisfies the following properties.

(i) Expectation of a constant is the constant itself i.e. $E(c) = c$, where c is a constant.

Proof:

By definition

$$E(X) = \Sigma x\, p(x)$$

As the variable X is taking only a constant value c therefore

$$E(c) = \Sigma c\, p(x)$$

$$= c\Sigma p(x)$$

$$= c\,(1), \text{ as sum of probabilities} = \Sigma p(x) = 1$$

$$= c$$

(ii) If a and b are two constants, then $E[aX + b] = a\,E(X) + b$

Proof:

By definition

$$E(X) = \Sigma x\, p(x)$$

Here we have a function of the random variable X i.e. $a\,X + b$, so

$$E(aX + b) = \Sigma(ax + b)\, p(x)$$

$$= \Sigma ax\, p(x) + \Sigma b\, p(x)$$

$$= a\Sigma x\, p(x) + b\Sigma p(x)$$

$$= a\,E(X) + b\,(1)$$

$$= a\,E(X) + b$$

This property shows that expectation is changed by that constant which is added to, subtracted from, multiplied with or divided by the values of a variable i.e.

(i) $E[X + c] = E(X) + c$ (ii) $E[X - c] = E(X) - c$

(iii) $E(cX) = c\,E(X)$ (iv) $E\left(\frac{X}{c}\right) = \frac{E(X)}{c}$

where X is a random variable and c is any constant.

Example 2.12

Suppose that X is a simple discrete random variable, distributed as follows:

X	$p(x)$
2	0.50
4	0.50

Find (i) $E(X)$ (ii) $E(X+10)$ (iii) $E(X - 10)$ (iv) $E(10X)$ (v) $E\left(\frac{X}{10}\right)$

Solution:

Given

X	$p(x)$	$X\,p(x)$
2	0.50	1
4	0.50	2
Total	1	3

(i) By definition $E(X) = \Sigma x\, p(x) = 3$

(ii) By property $E(X+10) = E(X) + 10 = 3 + 10 = 13$

(iii) $E(X-10) = E(X) - 10 = 3 - 10 = -7$

(iv) $E(10X) = 10E(X) = 10(3) = 30$

(v) $E\left(\frac{X}{10}\right) = \frac{E(X)}{10} = \frac{3}{10} = 0.3$

2.2.8 Variance and standard deviation of a random variable

If X is a random variable, then its variance is denoted by $Var(X)$ or $V(X)$ and its formula is given by

$$V(X) = E[X - E(X)]^2$$

Or $V(X) = E[X^2] - [E(X)]^2$

$$S.D(X) = \sqrt{V(X)}$$

Example 2.13

Consider the following discrete probability distribution:

X	:	3	8	10
$p(x)$	:	0.2	0.7	0.1

Determine the mean, variance and standard deviation.

Solution:

X	$p(x)$	$X\,p(x)$	$X^2\,p(x)$
3	0.2	0.6	1.8
8	0.7	5.6	44.8
10	0.1	1	10.0
Total	1	7.2	56.6

Mean $= E(X) = \Sigma x\, p(x) = 7.2$

Variance $= V(X) = E(X^2) - [E(X)]^2 = \Sigma x^2\, p(x) - [7.2]^2 = 56.6 - 51.84 = 4.76$

and $S.D(X) = \sqrt{V(X)} = \sqrt{4.76} = 2.181$

2.2.9 Properties of variance and standard deviation of a random variable

Following are some of the important properties.

(i) The variance and standard deviation of a constant is equal to zero i.e. $V(c) = 0$ and $S.D(c) = 0$, where c is a constant.

Proof:

By definition $V(X) = E[X - E(X)]^2$

As the variable X is taking only a constant value c, therefore

$$V(c) = E[c - E(c)]^2$$
$$= E[c - c]^2 \quad \because E(c) = c$$
$$= 0$$
$$S.D(c) = \sqrt{0} = 0$$

(ii) Variance and S.D of a random variable are not changed by adding/ subtracting a constant to/from the values of a random variable i.e.

$$V(X \pm c) = V(X)$$

Proof:

By definition $V(X) = E[X - E(X)]^2$

As $(X + c)$ is a function of the random variable X, therefore,

$$V(X+c) = E[X + c - E(X+c)]^2$$
$$= E[X + c - E(X) - E(c)]^2$$

$$= E[X + c - E(X) - c]^2 \quad \because E(c) = c$$

$$= E[X - E(X)]^2$$

$$= V(X)$$

Similarly $V(X - c) = V(X)$

$$\text{S.D}(X \pm c) = \text{S.D}(X)$$

(iii) If the values of a random variable are multiplied or divided by a constant, then its variance will change by the square of that constant i.e.

$$V(cX) = c^2 V(X)$$

$$V\left(\frac{X}{c}\right) = \frac{V(X)}{c^2}$$

Proof:

By definition $V(X) = E[X - E(X)]^2$

As (cX) is function of the random variable X so,

$$V(cX) = E[cX - E(cX)]^2$$

$$= E[cX - cE(X)]^2$$

$$= c^2 E[X - E(X)]^2$$

$$= c^2 V(X)$$

Similarly $V\left(\frac{X}{c}\right) = \frac{V(X)}{c^2}$

$$\text{S.D}(cX) = |c| \text{ S.D}(X)$$

$$\text{S.D}\left(\frac{X}{c}\right) = \frac{S.D(X)}{|c|}$$

(iv) Variance and standard deviation can never be negative i.e.
$V(X) \geq 0$ and $\text{S.D}(X) \geq 0$

Example 2.14

A random variable X has the probability distribution given below

X:	0	1	2	3
$p(x)$:	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

Find i) $E(X)$, ii) $V(X)$, iii) $V(X+5)$, iv) $V(X-3)$, v) $V(3X)$ vi) $\text{S.D}\left(\frac{X}{3}\right)$

Solution:

X	$p(x)$	$X\,p(x)$	$X^2\,p(x)$
0	3/10	0	0
1	4/10	4/10	4/10
2	2/10	4/10	8/10
3	1/10	3/10	9/10
Total	1	11/10	21/10

(i) $E(X) = \Sigma x\,p(x) = \frac{11}{10} = 1.1$

(ii) $V(X) = E(X^2) - [E(X)]^2 = \Sigma x^2\,p(x) - (1.1)^2$

$$= \frac{21}{10} - 1.21$$

$$= 2.1 - 1.21 = 0.89$$

(iii) By property of variance $V(X+5) = V(X) = 0.89$

(iv) $V(X-3) = V(X) = 0.89$

(v) $V(3X) = 3^2 V(X) = 9(0.89) = 8.01$

(vi) S.D $\frac{X}{3} = \frac{S.D(X)}{|3|}$ $\because$ S.D$(x) = \sqrt{0.89} = 0.94$

$$= \frac{0.94}{3} = 0.314$$

2.3 Probability distribution of a continuous random variable

All possible values of a continuous random variable along with probability cannot be presented in tabular form. This purpose is only achieved by formula, called probability density function of the continuous random variable.

2.3.1 Probability density function

If probability distribution of a continuous random variable is expressed by a formula, then it is called probability density function (pdf) or simply density function and is denoted by $f(x)$.

2.3.2 Properties of probability density function

A pdf must satisfy the following important properties:

(i) $f(x) \geq 0,\ a \leq x \leq b$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

Example: 2.15

A continuous random variable X has the probability function given by

$$f(x) = \begin{cases} cx, & 0 < x < 1 \\ 0, & otherwise \end{cases}$$

i. Find value of the constant c.
ii. Check that the function is a p.d.f

iii. Compute

(a) $P\left[\frac{1}{4} < X < \frac{1}{2}\right]$ (b) $P\left(X > \frac{1}{2}\right)$ (c) $P\left(X = \frac{1}{2}\right)$

Solution:

(i) We know that sum of all probabilities is equal to one i.e.

$$\int_0^1 f(x)\,dx = 1$$

$$\int_0^1 cx\,dx = 1$$

$$c\int_0^1 x dx = 1$$

$$c\left[\frac{x2}{2}\right]_0^1 = 1$$

$$c\left(\frac{1}{2} - 0\right) = 1$$

$$\frac{c}{2} = 1 \quad \Rightarrow c = 2$$

Put the value of c in the given function, we have

$$f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

(ii) a) $f(x) \geq 0$ for the given range

b) $\int_x f(x)dx = 1$

Taking L.H.S

$$\int_x f(x)dx = \int_0^1 2x\,dx = 2\int_0^1 x\,dx = 2\left|\frac{x^2}{2}\right|_0^1 = x^2\Big|_0^1 = 1 - 0 = 1 = R.H.S$$

Since both properties are satisfied, therefore, the given function is a p.d.f.

(iii) (a) $P\left(\frac{1}{4}<X<\frac{1}{2}\right)=\int_{1/4}^{1/2} f(x)\,dx=\int_{1/4}^{1/2} 2x\,dx=2\int_{1/4}^{1/2} x\,dx$

$$=2\left(\frac{x^2}{2}\right)\Big|_{1/4}^{1/2}=x^2\Big|_{1/4}^{1/2}=\frac{1}{4}-\frac{1}{16}=\frac{4-1}{16}=\frac{3}{16}$$

(b) $P\left(X>\frac{1}{2}\right)=\int_{\frac{1}{2}}^{1} f(x)\,dx=\int_{\frac{1}{2}}^{1} 2x\,dx=2\int_{\frac{1}{2}}^{1} x\,dx=2\frac{x^2}{2}\Big|_{1/2}^{1}$

$$=x^2\Big|_{1/2}^{1}=1-\frac{1}{4}=\frac{4-1}{4}=\frac{3}{4}$$

(c) $P\left(X=\frac{1}{2}\right)=0$, because in continuous case point probability is always equal to zero.

2.3.3 Expectation, variance and standard deviation of a continuous random variable

If X is a continuous random variable, then expected value of X is given by $E(X)=\int_x x f(x)\,dx$, provided that it exists.

Example 2.16

Let X be a continuous random variable with pdf given by

$$f(x)=\begin{cases}\frac{x}{2} & 0\le x\le 2\\ 0 & \textit{otherwise}\end{cases}.$$

Find the mean, variance and standard deviation of X.

Solution:

(i) By definition

$$\text{Mean}=E(X)=\int_x xf(x)dx$$

$$=\int_0^2 x\left(\frac{x}{2}\right)dx=\frac{1}{2}\int_0^2 x^2dx=\frac{1}{2}\left(\frac{x^3}{3}\Big|_0^2\right)$$

$$=\left(\frac{x^3}{6}\Big|_0^2=\frac{8}{6}-0=\frac{4}{3}=1.333\right)$$

(ii) By definition

$$V(X)=E(X^2)-[E(X)]^2$$

$$=\int_0^2 x^2f(x)\,dx-\left(\frac{4}{3}\right)^2$$

$$=\int_0^2 x^2\left(\frac{x}{2}\right)dx-\frac{16}{9}$$

$$=\int_0^2 \frac{x^3}{2}dx-\frac{16}{9}=\frac{x^4}{8}\Big|_0^2-\frac{16}{9}=\left(\frac{16}{8}-0\right)-\frac{16}{9}$$

$$=2-1.78=0.22$$

(iii) $\text{S.D}(X)=\sqrt{0.22}=0.47$

Example 2.17

A random variable X has the density function given by

$$f(x)=\begin{cases}(2-2x) & 0\le x\le 1\\ 0 & \textit{otherwise}\end{cases},$$

Find (i) $E(X)$ (ii) $E(X^2)$ (iii) $E(2X)$ (iv) $E(2X-1)$

(v) variance and standard deviation of X

Solution:

(i) $E(X) = \int_x x f(x)\,dx = \int_0^1 x(2-2x)\,dx = \int_0^1 (2x-2x^2)\,dx$

$= \int_0^1 2x\,dx - \int_0^1 2x^2\,dx = 2\int_0^1 x\,dx - 2\int_0^1 x^2 dx$

$= x^2\Big|_0^1 - 2\,\frac{x^3}{3}\Big|_0^1$

$= (1-0) - \left(\frac{2}{3}-0\right) = 1-\frac{2}{3} = \frac{3-2}{3} = \frac{1}{3}$

(ii) $E(X^2) = \int_x x^2 f(x)\,dx = \int_0^1 x^2(2-2x)\,dx = \int_0^1 (2x^2-2x^3)\,dx$

$= 2\int_0^1 x^2dx - 2\int_0^1 x^3dx = 2\left(\frac{x^3}{3}\Big|_0^1\right) - 2\left(\frac{x^4}{4}\Big|_0^1\right)$

$= 2\left(\frac{1}{3}-0\right) - 2\left(\frac{1}{4}-0\right) = \frac{2}{3} - \frac{1}{2} = \frac{4-3}{6} = \frac{1}{6}$

(iii) $E(2X) = 2E(X)$, (By property)

$= 2(1/3) = 2/3$

(iv) $E(2X-1) = 2E(X) - 1 = 2(\frac{1}{3}) - 1 = -1/3$

(v) Variance $= V(X) = E(X^2) - [E(X)]^2$

$= \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{6} - \frac{1}{9} = \frac{3-2}{18} = \frac{1}{18} = 0.056$

S.D$(X) = \sqrt{0.056} = 0.24$

2.4 Two independent random variables

This is an extension of the one random variable case. The problem will be to recognize the old ideas behind the new names.

2.4.1 Joint distributions

The distribution of two or more random variables is called joint distribution. Specifically;

The distribution of one random variable is called univariate probability distribution.

The distribution of two random variables is called bivariate probability distribution.

The distribution of three random variables is called trivariate probability distribution.

The distribution of many random variables is called multivariate probability distribution.

2.4.2 Bivariate probability distribution

If all possible values of two random variables along with their joint probabilities are presented in tabular form then, it is called bivariate probability distribution. Suppose X and Y are two discrete random variables, where X has "m" values and Y has "n" values, then bivariate probability distribution of X and Y is given by

X \ Y	y_1	y_2	...	y_j	...	y_n	$p(x)$
x_1	$p(x_1,y_1)$	$p(x_1,y_2)$	...	$p(x_1,y_j)$	...	$p(x_1,y_n)$	$p(x_1)$
x_2	$p(x_2,y_1)$	$p(x_2,y_2)$	...	$p(x_2,y_j)$	...	$p(x_2,y_n)$	$p(x_2)$
⋮			⋮		⋮	⋮	⋮
x_i	$p(x_i,y_1)$	$p(x_i,y_2)$	...	$p(x_i,y_j)$	...	$p(x_i,y_n)$	$p(x_i)$
⋮			⋮		⋮	⋮	⋮
x_m	$p(x_m,y_1)$	$p(x_m,y_2)$	...	$p(x_m,y_j)$	...	$p(x_m,y_n)$	$p(x_m)$
$p(y)$	$p(y_1)$	$p(y_2)$	...	$p(y_j)$	...	$p(y_n)$	1

♦ Marginal probability distributions

From the bivariate probability distribution we can obtain univariate probability distributions as:

(i) The marginal probability distribution of X

X	$p(x)$
x_1	$p(x_1)$
x_2	$p(x_2)$
⋮	⋮
x_i	$p(x_i)$
⋮	⋮
x_m	$p(x_m)$
Total	1

(ii) The marginal probability distribution of Y

Y	$p(y)$
y_1	$p(y_1)$
y_2	$p(y_2)$
⋮	⋮
y_j	$p(y_j)$
⋮	⋮
y_n	$p(y_n)$
Total	1

2.4.3 Bivariate probability function

If all possible values of two random variables with their associated probabilities are shown by a formula, then it is called bivariate probability function. It is denoted by $p(x_i, y_j)$ (discrete case) and $f(x_i, y_j)$ (continuous case) ; for all x_i and y_j.

2.4.4 Independence of two random variables

Two random variables X and Y are statistically independent if and only if their bivariate probability function can be expressed as the product of the marginal probability functions i.e.

$$f(x_i, y_j) = f(x_i) f(y_j)$$

2.4.5 Mathematical expectation of two random variables

If X and Y are two random variables then mean or expected value of their sum or difference is denoted by $E(X \pm Y)$ and is defined as:

$$E(X \pm Y) = \sum_i \sum_j (x_i \pm y_j)\ p(x_i, y_j) \qquad \text{(discrete random variables case)}$$

$$= \iint_{x\,y} (x_i \pm y_j)\ f(x_i, y_j)\ d_{xi}\ d_{yj} \qquad \text{(continuous random variables case)}$$

Expected value of their product is

$$E(XY) = \sum_j \sum_i (x_i y_j)\ p(x_i, y_j) \qquad \text{(for discrete case)}$$

$$= \iint_{x\,y} (x_i y_j)\ f(x_i, y_j)\ d_{xi}\ d_{yj} \qquad \text{(for continuous case)}$$

2.4.6 Properties of mathematical expectation of two random variables

Let X and Y are two independent random variables and then they jointly have the following properties:

(i) The expected value of the sum of two random variables is equal to the sum of the expected values of the individual random variables i.e.

$$E(X+Y) = E(X) + E(Y)$$

Proof:

By definition

$$E(X+Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_i + y_j)\ p(x_i, y_j)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} x_i\ p(x_i, y_j) + \sum_{i=1}^{m} \sum_{j=1}^{n} y_j\ p(x_i, y_j) \qquad (1)$$

Consider $\sum_{i=1}^{m} \sum_{j=1}^{n} x_i\ p(x_i, y_j)$

$$= \sum_{i=1}^{m} x_i \sum_{j=1}^{n} p(x_i, y_j)$$

$$= \sum_{i=1}^{m} x_i [p(x_i, y_1) + p(x_i, y_2) + \ldots + p(x_i, y_n)]$$

$$= \sum_{i=1}^{m} x_i [p(x_i)\ p(y_1) + p(x_i)\ p(y_2) + \ldots + p(x_i)\ p(y_n)]$$ (as X and Y are independent)

$$= \sum_{i=1}^{m} x_i\, p(x_i) [p(y_1) + p(y_2) + \ldots + p(y_n)]$$

$$= \sum_{i=1}^{m} x_i\, p(x_i)\ (1)$$ (as sum of probabilities is equal to one.)

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

$$= E(X) \qquad (2)$$

Again consider $\sum_{i=1}^{m}\sum_{j=1}^{n} y_j\, p(x_i, y_j)$

$$= \sum_{i=1}^{n} y_j \sum_{j=1}^{m} p(x_i, y_j)$$

$$= \sum_{j=1}^{n} y_j [p(x_i, y_j) + p\ x_2, y_j) + \ldots + p(x_m, y_j)]$$

$$= \sum_{j=1}^{n} y_j [p(x_1)\, p(y_j) + p(x_2)\, p(y_j) + \ldots + p(x_m) p(y_j)]$$

$$= \sum_{j=1}^{n} y_j\, p(y_j) [p(x_1) + (p\ x_2) + \ldots + p(x_m)]$$

$$= \sum_{j=1}^{n} y_j p(y_j)\ (1)$$

$$= \sum_{j=1}^{n} y_j\, p(y_j)$$

$$= E(Y) \qquad (3)$$

Put equation (2) and (3) in equation (1), we get $E(X+Y) = E(X) + E(Y)$

(ii) The expected value of the product of two independent random variables is equal to the product of their individual expected values i.e. $E(X\,Y) = E(X)\,E(Y)$

Proof:

By definition

$$E(X\,Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} (x_i, y_j)\, p(x_i, y_j)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} x_i\, y_j\, p(x_i)\, p(y_j)$$ (as X and Y are independent)

$$= \sum_{i=1}^{m} x_i\, p(x_i) \sum_{j=1}^{n} y_j\, p(y_j)$$

$$= E(X)\, E(Y)$$

Example 2.18

Let X and Y are two discrete random variables with the following joint probability distribution:

X \ Y	1	3	5
2	0.10	0.20	0.10
4	0.15	0.30	0.15

Compute; E(X), E(Y), E($X+Y$), E ($2X-3Y$) and E($X\,Y$).

Solution:

Adding rows and columns to find $p(x)$ and $p(y)$ shown in the following table.

X \ Y	1	3	5	$p(x)$
2	0.10	0.20	0.10	0.40
4	0.15	0.30	0.15	0.60
$p(y)$	0.25	0.50	0.25	1

By definition

$$E(X) = \Sigma x\, p(x) = 2 \times 0.40 + 4 \times 0.60 = 0.80 + 2.40 = 3.2$$

$$E(Y) = \Sigma y\, p(y) = 1 \times 0.25 + 3 \times 0.50 + 5 \times 0.25$$

$$= 0.25 + 1.50 + 1.25 = 3$$

Now by property

$$E(X + Y) = E(X) + E(Y) = 3.2 + 3 = 6.8$$

$$E(2X - 3Y) = E(2X) - E(3Y)$$

$$= 2E(X) - 3E(Y)$$

$$= 2(3.2) - 3(3) = 6.4 - 9 = -2.6$$

Since X and Y are independent, therefore,

$$E(X\,Y) = E(X)\,E(Y) = (3.2)\,(3) = 9.6$$

2.4.7 Variance of the sum or difference of two independent random variables

If X and Y are two independent random variables, then variance of their sum is given by the formula:

$$V(X+Y) = E[X+Y - E(X+Y)]^2$$

$$= E[X+Y - E(X) - E(Y)]^2$$

$$= E[\{X - E(X)\} + \{Y - E(Y)\}]^2$$

$$= E[\{X - E(X)\}^2 + \{Y - E(Y)\}^2 + 2\{X - E(X)\}\{Y - E(Y)\}]$$

$$= E[X - E(X)]^2 + E[Y - E(Y)]^2 + 2\,E\{X - E(X)\}\{Y - E(Y)\}$$

$$= V(X) + V(Y) + 2\,\text{cov}(X, Y)$$

When X and Y are independent then $\text{cov}(X, Y) = 0$

$\therefore$ $V(X+Y) = V(X) + V(Y)$, Similarly

$$V(X - Y) = V(X) + V(Y)$$

Example 2.19

Consider the following joint probability distribution of two independent random variables X and Y.

X \ Y	0	1	2
0	0.10	.20	.10
1	0.15	.30	.15

Find: (i) $V(X)$, (ii) $V(Y)$, (iii) $V(X+Y)$, (iv) $V(X-Y)$

Solution:

X \ Y	0	1	2	$p(x)$
0	0.10	.20	.10	.40
1	0.15	.30	.15	.60
$p(y)$	0.25	.50	.25	1

(i) $V(X) = E(X)^2 - [E(X)]^2$

$$E(X) = \Sigma x\, p(x) = 0\,(.40) + 1\,(.60) = 0 + .60 = 0.60$$

$$E(X^2) = \Sigma x^2\, p(x) = 0^2\,(.40) + 1^2\,(.60) = 0 + .60 = 0.60$$

$$\therefore\ V(X) = 0.60 - (0.60)^2 = 0.60 - 0.36 = 0.\tilde{2}4$$

(ii) $V(Y) = E(Y^2) - [E(Y)]^2$

$$E(Y) = \Sigma y\, p(y) = 0(.25) + 1(.50) + 2(.25) = 0 + 0.50 + 0.5 = 1$$

$$E(Y^2) = \Sigma y^2\, p(y) = 0^2\,(.25) + 1^2\,(.50) + 2^2\,(.25) = 0 + .50 + 1 = 1.50$$

$$\therefore\ V(Y) = 1.50 - (1)^2 = 1.50 - 1 = 0.50$$

Since X and Y are independent, therefore,

(iii) $V(X+Y) = V(X) + V(Y) = 0.24 + 0.50 = 0.74$

(iv) $V(X-Y) = V(X) + V(Y) = 0.24 + 0.50 = 0.74$

Key points

- A variable that is itself a function of the results of a random experiment is called random variable
- A variable which takes jumping values or isolated values is called discrete random variable
- A variable which takes any value between two limits $[a, b]$, $a < b$ is called continuous random variable
- If all possible values of a random variable along with their respective probabilities are shown in tabular form and sum of probabilities is equal to one, then such tabular form is called probability distribution
- If all possible values of a random variable along with their respective probabilities are shown by a formula, then it is called probability function or probability mass function. It is denoted by $p(x_i)$
- The mean of a random variable is called mathematical expectation.
- $E(c) = c$,where c is a constant
- Mean : $E(X) = \Sigma x\, p(x)$ where X is a discrete random variable
- Variance : $V(X) = E[X^2] - [E(X)]^2$
- The distribution of two or more random variables is called joint distribution.
- If $f(x_i, y_j) = f(x_i)\, f(y_j)$, then X and Y are independent
- $E(X+Y) = E(X) + E(Y)$ and $E(X-Y) = E(X) - E(Y)$
- $E(X\,Y) = E(X)\,E(Y)$, if X and Y are independent.
- $V(X+Y) = V(X) + V(Y)$ and $V(X-Y) = V(X) + V(Y)$, if X and Y are independent.

Exercise

2.1 Read the following statements carefully and write T for true and F for false statement.

i. Random variable is also called chance variable.
ii. Discrete random variable takes every value between two limits
iii. The sum of probabilities in a probability distribution must be one.
iv. The probability function and pdf can be negative.
v. Mathematical expectation of a random variable is also called average or mean of the random variable.
vi. Expected value of a constant is zero.
vii. Variance of a constant is zero.
viii. $E(X-Y) = E(X) - E(Y)$
ix. $V(X-Y) = V(X) - V(Y)$
x. If $f(x, y) = f(x)\, f(y)$, it means X and Y are independent.

2.2 Fill in the blanks.

i. A discrete variable can take a _____ number of values within its range or an infinite number of values that are countable.
ii. The probability function $p(x)$ cannot exceed_____
iii. if $p(x) = \frac{x}{21}$ for $x = 1, 2, 3, 4, 5, 6$ then $P(X = 2 \text{ or } 3) =$_____
iv. A variable whose value are obtained from the outcomes of a random experiment is called _____
v. Random variables are classified into _____
vi. The distribution of two or more random variables is called_____
vii. The value of the expression $\sum_i \sum_j p(x_i, y_j)$ is always_____
viii. If X and Y are two independent variables then $E(X\,Y) =$_____
ix. If X and Y are independent random variables then $S.D(X-Y) =$____
x. If $E(X) = \frac{2}{3}$, $E(X^2) = \frac{8}{9}$ then $S.D(X) =$ _____

2.3 Select the correct answer out of the given ones.

(i) The height of persons in a country is

(a) discrete random variable

(b) continuous random variable

(c) both discrete and continuous

(d) neither discrete nor continuous

(ii) The outcomes of tossing a coin three times is a variable of the type

(a) continuous (b) discrete

(c) neither discrete not continuous

(d) both discrete and continuous

(iii) If $f(x) = kx$, $0 < x < 1$, then value of k is

(a) $\frac{1}{3}$ (b) 2 (c) $\frac{1}{2}$ (d) 1

(iv) If $V(X) = 2$ then $V(2X + 5)$ is equal to

(a) 4 (b) 2 (c) 8 (d) 6

(v) If $E(X) = 4$ then $E[3X + 10]$ is

(a) 10 (b) 13 (c) 4 (d) 22

(vi) If $E(X) = \frac{2}{3}$, $E(X^2) = \frac{8}{9}$ then S.D of X is

(a) $\frac{4}{9}$ (b) $\frac{9}{4}$ (c) $\frac{2}{3}$ (d) $\frac{2}{9}$

(vii) If $f(x, y) = f(x)\, f(y)$, it means variables are

(a) correlated (b) dependent

(c) independent (d) associated

(viii) If X and Y are independent random variables then $V(X-Y) =$

(a) $V(X) - V(Y)$ (b) $V(X) + V(Y)$

(c) $V(X)\, V(Y)$ (d) $V(X + Y)$

(ix) continuous probability distributions give

(a) interval probability (b) point probability

(c) negative probability (d) zero probability

(x) if X is a random variable having its pdf, the $E(X)$ is called

(a) median (b) geometric mean

(c) mode (d) arithmetic mean

2.4 Describe the concept of a random variable and give its examples.

2.5 Explain different types of a random variable with examples.

2.6 Classify each of the following random variables as either discrete or continuous:

i. Time to failure for an electronic system.

ii. The height of a person.

iii. The number of questions asked in an oral examination.

iv. Temperature at a place.

v. The maximum breaking strength 250 kg of a wire.

vi. The number of fatal traffic accident per month on the motor way.

vii. The life time of a mobile set.

viii. The amount of rainfall at Islamabad during different months of 2016.

ix. The number of admitted patients in a hospital in a year.

x. The number of Mosque per village in a district.

2.7 Define probability distribution, probability function and probability density function. What are the two basic properties of all probability functions?

2.8 Find probability distribution for the number of heads when 2 balanced coins are tossed.

2.9 In a family of three children find the probability distribution for the number of girls.

2.10 A random variable has the following probability distribution.

X	4	6	7	10
$p(x)$	0.2	0.4	c	0.1

(i) Find the value of c. (ii) $P(X < 7)$ (iii) $P(X \geq 6)$

2.11 A coin is tossed three times in succession. Find the probability distribution for the number of heads minus number of tails.

2.12 A pair of fair dice is rolled once. Find the probability distribution for the difference of dots.

2.13 Check whether the following is a probability distribution or not? If not, then why?

X	0	10	15	25	50
$p(x)$	0.1	0.3	0.4	0.3	0.1

2.14 A discrete random variable has the probability mass function:

$$p(x) = \begin{cases} \frac{^3C_x}{8} & , \quad x = 0,1,2,3 \\ 0 & , \quad elsewhere \end{cases}$$

Find (i) $P(X = 2)$ (ii) $P(X > 1)$ (iii) $P(0 < X < 3)$

2.15 (a) Define p.d.f and state its properties.

(b) The density function of a continuous random variable X is:

$$f(x) = \begin{bmatrix} Ax^2 & , & 0 \leq x \leq 1 \\ 0 & , & elsewhere \end{bmatrix}$$

Find (i) value of the constant A (ii) $P(X \leq 0.5)$ (ii) $P(0.2 < X < 0.3)$.

2.16 Discuss mathematical expectation. Write some of its important properties.

2.17 For the following discrete probability distribution of X

X	-2	1	2
$p(x)$	1/3	1/6	1/2

Find (i) $E(X)$ (ii) $E(X^2)$ (iii) $E(2X + 5)$

2.18 Find the mean, variance and standard deviation for the following probability distribution:

Y	-1	0	1	2	3
$p(y)$	0.125	0.5	0.2	0.05	0.125

2.19 Let Y be a random variable with probability distribution as follows:

Y	1	2	3	4	5
$p(y)$	0.125	0.45	0.25	0.05	0.125

Find (i) Expected value

(ii) Variance

(iii) S.D for the random variable Y

2.20 A variable X has the p.d.f $f(x) = \begin{bmatrix} kx^3(1-x) & , & 0 \leq x \leq 1 \\ 0 & , & otherwise \end{bmatrix}$

Find (i) The value of k (ii) $E(X)$ (iii) Variance (iv) S.D

2.21 Find mean, variance and standard deviation for the random variable Y whose p.d.f is $f(y) = \begin{cases} \frac{2}{3}(2-y) & , \quad 0 \le y \le 1 \\ 0 & , \quad elsewhere \end{cases}$

2.22 The joint distribution of two independent random variables X and Y is

X \ Y	0	1
1	1/6	1/6
2	1/6	1/6
3	1/6	1/6

Find $E(X)$, $E(Y)$, $E(X+Y)$, $E(XY)$

2.23 Suppose that X and Y have the following joint distribution.

X \ Y	10	20	40	80
20	0.04	0.08	0.08	0.05
40	0.12	0.24	0.24	0.15

(i) Are X and Y independent?

(ii) Find $E(X)$, $E(Y)$, $E(X+Y)$, $E(XY)$

2.24 For the following joint distribution:

X \ Y	1	2
0	.06	.04
1	.30	.20
2	.24	.16

Find (i) $V(X)$ (ii) $V(Y)$

(iii) $V(X+Y)$ (iv) $V(X-Y)$

Unit -03

Special Discrete Probability Distributions

After studying this unit, the students will be able to

- Define discrete uniform random variable, discrete uniform probability distribution and discrete uniform probability mass function.
- Calculate mean, variance and standard deviation of discrete uniform probability distribution
- Define random digits/numbers and to know, how the random digits/numbers are generated.
- Solve real life problems using discrete uniform probability distribution.
- Define Bernoulli trials, Bernoulli probability distribution and Bernoulli mass function.
- Calculate mean, variance and standard deviation of Bernoulli probability distribution.
- Solve real life problems using Bernoulli probability distribution.
- Define Binomial experiment, Binomial random variable, Binomial probability distribution, Binomial probability mass function and Binomial frequency distribution.
- Calculate mean, variance and standard deviation of Binomial probability distribution
- Solve real life problems using Binomial probability distribution.
- Define hypergeometric experiment, hypergeometric random variable, hypergeometric probability distribution and hypergeometric probability mass function.
- Calculate mean, variance and standard deviation of hypergeometric probability distribution
- Solve real life problems using hypergeometric probability distribution.

3.1 When to use discrete probability distributions

Discrete probability distributions are used to compute probabilities for all possible values of discrete random variables, by an easy way. In this unit a few simple and commonly used distributions are given that cover many areas in our surrounding. It is important to know and remember that which probability distribution is suitable for a particular situation.

3.1.1 Introduction to discrete uniform distribution

This is the simplest probability distribution among the discrete probability distributions. It is used in the experiments/situations where probability at every point remains the same e.g. the outcomes of rolling a fair die, drawing cards from a well shuffled deck of cards, drawing of prize bond number etc. follow uniform distribution. A variable denoting the outcomes of such uniform experiments is called discrete uniform random variable and its probability distribution is called discrete uniform probability distribution. The most important application of the uniform distribution is in the generation of random numbers.

3.1.2 Definition

A probability distribution of the type:

X	x_1	x_2	x_3	...	x_N	Total
$p(x)$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	1

is called discrete uniform probability distribution.

Example 3.1

Find probability distribution for the outcomes of a fair die when it is rolled once.

Solution:

When a fair die is rolled once, then all outcomes are equally likely, therefore, its probability distribution is given by

X	1	2	3	4	5	6	Total
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Note that X follows discrete uniform distribution.

3.1.3 Definition of discrete uniform probability function

A probability function defined as $p(x)=\begin{cases}\frac{1}{N} & , \quad x=1,2,\cdots,N \\ 0 & , \quad otherwise\end{cases}$

is called discrete uniform probability function and the variable X is called discrete uniform random variable. This distribution has only one parameter "N", the total number of results /items of a uniform experiment.

Example 3.2

Write discrete uniform probability function for the results of a fair die when it is rolled once.

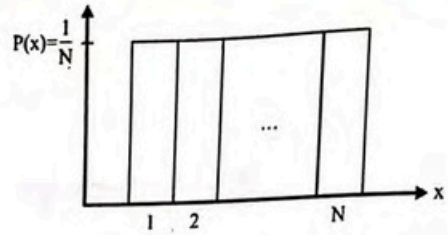
Solution:

Here random variable X takes all of its values with equal probability $\frac{1}{6}$.

Thus, discrete uniform probability function will be of the form:

$$p(x)=\begin{bmatrix}\frac{1}{6} & , \quad x=1,2,3,4,5,6 \\ 0 & , \quad otherwise\end{bmatrix}$$

♦ Graphical representation of discrete uniform distribution by means of a histogram is a set of adjacent rectangles with equal heights i.e.



Due to rectangular shape, uniform distribution is also known as rectangular distribution

3.1.4 Properties of discrete uniform probability distribution

(i) Mean of discrete uniform distribution

By definition mean $= E(X) = \sum_x x\,p(x)$

Put range and formula of discrete uniform distribution, we get

$$E(X) = \sum_{x=1}^{N} x\frac{1}{N} = \frac{1}{N}\sum_{x=1}^{N} x = \frac{1}{N}[1+2+3+\cdots+N]$$

$$= \frac{1}{N}\left[\frac{N(N+1)}{2}\right] \quad as\ 1+2+3+\cdots+N = \frac{N(N+1)}{2}$$

$$= \frac{N+1}{2}$$

(ii) Variance of discrete uniform distribution

By definition variance: $V(X) = E(X^2) - [E(X)]^2$ (1)

We know that $E(X) = \frac{N+1}{2}$ (2)

Now $E(X^2) = \sum_{x=1}^{N} x^2\frac{1}{N} = \frac{1}{N}\sum_{x=1}^{N} x^2 = \frac{1}{N}[1^2+2^2+\cdots+N^2]$

$$= \frac{1}{N}\left[\frac{N(N+1)(2N+1)}{6}\right], \quad as\ 1^2+2^2+\ +N^2 = \frac{N(N+1)(2N+1)}{6}$$

$$= \frac{(N+1)(2N+1)}{6} \qquad (3)$$

Put equation (2) and (3) in equation (1) we get,

$$V(X) = \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{N+1}{2}\left[\frac{2N+1}{3} - \frac{N+1}{2}\right]$$

$$= \frac{N+1}{2} - \left[\frac{4N+2-3N-3}{6}\right] = \frac{N+1}{2}\left[\frac{N-1}{6}\right]$$

$$= \frac{N^2-1}{12}$$

(iii) Standard deviation of uniform distribution

By definition S.D$(X) = \sqrt{V(X)} = \sqrt{\frac{N^2-1}{12}}$

Example 3.3

If X is a discrete uniform random variable taking values 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 each with probability equal to $\frac{1}{10}$. Find its mean, variance and standard deviation.

Solution:

Here total number of observations $= N = 10$

$$\therefore \text{Mean} = \frac{N+1}{2} = \frac{10+1}{2} = \frac{11}{2} = 5.5$$

$$\text{Variance} = \frac{N^2-1}{12} = \frac{10^2-1}{12} = \frac{100-1}{12} = \frac{99}{12} = 8.25$$

$$\text{Standard deviation} = \sqrt{8.25} = 2.87$$

3.1.5 Random digits/numbers

There are ten digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Involving equal chance $\frac{1}{10}$ of selection for each digit, then they are called random digits because now personal liking or disliking is impossible or one cannot predict in advance that which digit will come out during the selection process. Similarly pair of digits make a number e.g. 00, 01, 02,..., 99 each having probability $\frac{1}{100}$ or 001, 002, 003, ..., 1000 each having probability $\frac{1}{1000}$ and so on, are called random numbers. Thus, random digits/numbers can simply be defined as the digits or numbers which have equal chances of occurrence or their selection cannot be predicted in advance are called random digits/numbers.

3.1.6 Random number generator

Random number generator is a process (computational or physical device) designed to generate a sequence of numbers that cannot be accurately predicted. One of the simple procedures is that: write each digit separately on a slip of paper of equal size, fold it and put them in a basket or a bowl etc. and mix them thoroughly. Then draw out the digits blindly one by one with replacement and write in the order they occur and group them for convenience in blocks of two, three, four etc. The resulting set of random digits is called random number table. Nowadays several random number tables exist constructed by Tippet, Kendall and Smith, Fisher and Yates, RAND Corporation etc. Random numbers are also available in calculators, computer packages and on the Internet. In the real life, these random digits/numbers are used to make the selection process fair and free from personal biases. These are extensively used in the random sampling for selection of a random sample, which everywhere is necessary for precise statistical inference.

Table 3.1 Random digits (blocked merely for convenience)

3 6 7 4 4	1 9 6 6 6	2 2 3 3 6	5 2 9 1 9	9 0 7 9 0
9 5 6 5 5	9 0 9 4 1	0 6 6 1 2	8 4 7 4 7	5 6 0 9 0
7 7 2 7 9	6 9 6 1 1	3 5 3 1 1	4 8 7 5 3	3 6 3 2 4
3 1 3 0 0	5 7 0 2 1	1 6 4 9 9	7 3 5 1 3	0 2 8 0 6
7 2 4 3 8	5 6 4 9 1	9 4 0 1 6	2 0 0 5 8	5 7 1 5 4
2 0 7 3 4	1 7 7 7 9	8 0 7 7 6	3 5 9 1 0	9 8 1 2 9
7 1 2 6 1	1 9 2 7 5	2 3 4 7 4	6 8 8 0 2	9 2 7 4 8
5 7 5 9 7	7 5 1 8 8	4 3 5 7 8	9 1 4 9 9	3 2 0 5 0
3 4 7 8 7	6 5 0 3 3	0 3 6 1 2	3 0 7 7 5	2 9 5 0 4
7 8 9 8 4	3 2 6 4 0	1 1 0 0 7	5 7 9 1 3	8 9 2 1 1
0 8 0 1 9	8 5 8 2 9	5 2 0 2 4	2 7 5 0 8	7 6 8 7 6
2 9 8 6 4	5 3 3 9 5	6 7 9 4 3	1 8 5 9 2	2 1 8 3 1
8 1 1 7 0	3 7 4 1 3	4 1 9 5 9	4 9 4 7 5	0 1 2 5 9
7 8 5 0 7	7 9 9 2 8	8 3 3 5 6	1 2 5 1 1	9 8 5 8 4
9 8 7 7 1	8 0 3 5 2	0 0 4 4 2	1 3 5 2 4	5 4 1 2 8
8 3 1 0 5	9 9 9 9 4	0 6 1 1 0	4 6 9 5 7	4 5 7 4 9
1 0 5 0 0	7 6 1 4 6	5 7 1 1 8	5 5 1 5 2	6 0 1 7 8
0 8 8 7 4	6 2 6 8 8	8 6 7 4 6	9 3 1 2 1	6 4 8 2 7
4 9 5 3 3	7 8 1 5 4	2 5 1 2 9	2 0 6 5 4	9 2 0 8 4
7 0 6 7 1	1 2 3 0 1	7 5 0 4 2	8 4 7 3 4	5 3 0 4 7
9 0 3 0 0	3 1 0 5 3	7 1 3 3 2	6 9 2 9 3	3 6 7 3 3
3 5 1 3 7	4 8 4 2 5	4 3 9 5 2	8 5 3 2 5	6 3 0 5 3
2 8 1 4 9	1 2 9 4 1	7 5 9 8 8	1 7 9 3 5	9 1 0 4 6
1 9 1 7 9	1 0 9 5 8	6 1 4 4 6	3 9 3 7 5	8 6 4 1 7
9 1 9 0 9	2 3 8 2 7	7 5 9 1 9	0 6 8 2 2	5 2 1 3 4
5 8 4 6 7	7 7 3 8 1	9 7 5 3 1	9 1 7 5 1	6 0 1 2 4
9 0 3 5 7	1 6 8 9 5	7 3 6 9 4	2 4 9 3 6	2 9 7 7 6
7 8 1 5 3	0 5 1 2 9	7 1 1 5 6	0 4 0 2 4	6 9 6 5 3
6 2 8 5 1	5 7 7 1 3	4 3 7 9 2	7 3 3 0 0	6 8 2 2 2
9 6 8 6 3	9 1 4 7 2	8 8 5 3 9	3 7 2 4 5	0 2 9 0 5
4 4 1 2 0	8 6 9 9 1	8 8 8 0 4	7 4 8 5 5	7 8 0 9 3
1 7 0 5 3	7 3 3 5 7	1 3 3 4 9	7 5 5 0 1	9 8 1 7 0
9 9 1 6 6	0 6 7 5 6	9 7 4 8 7	2 1 9 5 2	1 8 1 4 2
1 4 4 3 2	8 1 4 1 9	2 9 3 9 9	9 8 4 1 3	4 5 1 7 3
8 0 5 1 4	0 5 8 0 4	4 4 3 9 2	7 6 7 0 8	5 0 4 9 3
0 6 4 8 7	8 2 5 8 0	8 0 5 4 2	2 5 1 8 6	0 3 2 9 6
6 7 7 6 9	8 8 8 4 0	6 7 0 2 6	8 3 2 7 0	6 6 7 2 1
7 2 7 3 9	9 5 4 6 6	4 1 6 1 6	9 7 0 0 1	1 5 0 2 2
5 4 2 1 7	4 2 7 2 1	0 9 7 1 0	8 3 1 3 1	8 8 1 1 3
9 0 4 3 5	2 9 2 3 9	6 4 6 0 8	1 0 5 9 4	1 3 7 6 3

6 6 0 3 4	7 5 3 0 9	9 6 0 4 6	2 2 1 6 8	9 9 4 4 4
3 2 6 4 1	9 3 6 2 5	4 1 9 3 2	0 1 4 8 6	4 5 8 6 5
7 4 2 5 9	7 9 9 3 6	3 8 5 0 0	8 4 0 7 7	1 9 8 8 4
8 7 3 3 0	9 3 6 8 3	4 5 2 5 9	5 3 1 2 3	4 3 7 1 0
8 6 6 1 4	9 4 7 6 4	5 1 7 1 8	5 9 5 9 5	2 6 2 4 3
7 8 2 5 3	7 8 2 6 8	3 6 1 3 1	9 7 0 9 2	4 2 0 2 1
4 6 9 1 7	2 9 0 7 5	8 4 3 0 5	7 6 4 2 3	7 7 1 0 2
7 0 2 0 7	6 7 5 3 1	8 6 8 3 8	2 8 9 9 1	5 0 6 8 4
5 8 8 5 4	0 8 3 7 4	6 8 6 5 3	8 0 8 2 8	8 8 5 3 3
6 8 7 9 1	6 7 7 8 8	5 8 9 8 9	8 2 4 7 3	5 1 6 9 8
2 1 6 6 8	8 0 2 5 5	6 6 8 9 4	1 2 0 9 8	4 3 3 8 4
2 7 8 5 4	7 2 2 7 1	8 9 0 5 4	1 9 1 5 0	9 4 5 6 7
1 3 2 5 4	3 7 3 7 0	7 5 9 3 5	8 7 3 8 1	0 5 9 3 7
9 6 7 9 6	9 7 2 7 9	9 7 2 6 9	9 4 9 2 5	8 8 4 4 4
1 7 2 0 7	2 0 2 2 9	2 4 4 8 1	8 8 6 6 3	0 7 9 7 2
6 7 3 2 7	8 6 4 5 3	2 5 4 4 1	7 0 1 5 1	9 1 1 4 5
7 4 7 7 6	9 6 6 9 6	8 4 6 9 8	5 7 1 8 2	5 3 6 9 2
8 3 6 1 1	7 7 3 9 1	0 5 7 3 2	9 3 9 5 3	3 3 5 7 1
0 2 2 2 0	6 3 1 0 0	5 5 4 7 9	6 5 4 6 7	4 5 2 1 4
3 8 8 6 8	9 0 6 9 5	3 8 7 0 7	6 6 5 5 9	5 6 0 9 7
0 3 1 2 5	3 7 9 1 8	2 7 3 2 6	5 7 5 6 9	6 6 9 2 6
4 1 7 1 6	3 3 9 9 7	6 2 9 7 7	3 7 7 8 3	0 1 7 2 1
6 0 9 0 9	8 1 7 4 9	4 0 6 9 5	4 6 3 0 0	9 9 4 5 4
1 6 8 3 1	7 4 7 3 6	3 0 5 8 0	5 0 3 1 7	8 9 6 0 7
2 6 3 2 0	9 5 9 6 2	5 7 0 4 3	5 6 0 6 8	9 9 1 7 8
3 8 5 6 0	9 3 3 1 8	2 0 5 8 4	6 5 5 1 6	0 2 0 0 0
1 3 2 7 9	9 8 5 7 6	1 3 5 0 2	2 8 2 0 4	8 9 9 2 0
5 9 5 0 9	3 6 2 7 5	5 3 9 5 8	2 7 6 7 7	6 6 8 9 6
5 7 9 3 2	1 4 2 3 2	8 1 9 2 4	4 1 0 7 4	7 1 3 2 0
8 1 6 0 4	8 6 3 4 7	5 3 9 4 4	9 8 9 9 9	4 6 2 3 9
9 2 5 6 3	0 1 6 7 3	8 0 0 7 8	0 1 1 8 5	5 3 5 7 6
3 2 3 4 9	7 1 3 6 5	7 3 4 9 8	8 3 3 5 1	5 4 7 2 9
7 7 5 5 7	9 3 7 9 9	5 0 7 7 4	8 6 1 2 4	6 6 9 1 0
8 6 8 4 4	2 8 0 6 2	2 6 9 9 5	2 3 8 7 4	9 6 2 9 9
6 8 3 9 8	0 5 2 6 5	4 7 6 1 7	5 6 1 6 1	0 9 5 6 5
1 1 1 6 2	0 7 5 0 9	6 2 0 8 7	5 6 2 2 1	8 9 5 4 7
4 8 0 1 0	2 9 6 6 2	4 2 6 8 3	2 8 3 2 6	1 9 3 3 1
2 8 7 0 5	4 8 5 3 1	7 1 1 8 2	7 0 0 1 0	0 2 5 6 2
7 9 3 0 3	2 1 9 0 2	1 7 7 9 1	4 9 5 9 8	9 8 5 1 9
7 4 0 5 9	8 0 9 0 7	2 3 3 9 2	9 9 7 4 2	6 8 7 7 1

3.1.7 Procedure of selecting random sample using random numbers

(i) If a population has $N = 10$ elements, allot digits 0, 1, 2....9 to elements of the population. If $N = 100$, then numbers 00, 01, 02... 99 are allotted. If $N = 1000$, then numbers 001, 002, 003... 1000 are allotted and so on.

(ii) Take any table of random numbers and take a random start from anywhere either from the top or bottom, vertically, horizontally or diagonally.

(iii) Draw a digit or number of two digits, number of three digits and so on according to your population size, ignore the number which is greater than the population size N or sampling is done without replacement and go further.

(iv) Continue the process till the desired number of sample units is obtained.

Example 3.4

Select a random sample of six students from a statistics class of 40 students by using a random number table.

Solution:

(i) Allot serial number 00, 01, 02, ..., 39 to students of the class.

(ii) Take a random number table and take numbers of two digits from anywhere, ignore the number which is greater than 39, and go further. Let the random numbers be 39, 37, 02, 10, 21, and 22.

(iii) The students whose serial numbers are 02, 10, 21, 22, 37, 39 are included in our sample. This is random sample because our personal judgment is not involved here and all the students were given equal opportunity through random number table for being included in the sample.

3.2 Bernoulli trail

A random experiment which has two possible outcomes classified as success and failure is called a Bernoulli trial. For example, gender of a child (male or female), performance of a student in an examination (pass or fail), result of tossing a coin (head or tail) etc. The random variable is taking only two values i.e. "0" for failure and "1" for success. Further, probability of success is denoted by p and failure by q such that $p+q=1$.

3.2.1 Bernoulli probability distribution

A probability distribution of the type:

X	$p(x)$
0	q
1	p
Total	1

is called Bernoulli probability distribution and the variable X which denotes the results of a Bernoulli trail is called Bernoulli random variable. This distribution was introduced by a Swiss mathematician Jacob Bernoulli (1654–1705). It is also known as point Binomial distribution because it has only two classes of events. Mathematically

$$p(x)=\begin{cases} p^x q^{1-x} & , \ x=0,1 \\ 0 & , \ otherwise \end{cases}$$

is called Bernoulli probability mass function. This distribution has only one parameter i.e. p.

Example 3.5

A fair coin is tossed o ce. What is the probability that (i) no head occurs? (ii) one head occurs?

Solution:

Tossing a coin is Bernoulli experiment. Let us define a random variable as

X: number of heads

$=0, 1$

p = probability of head $=\frac{1}{2}$ (as coin is fair).

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

The Bernoulli probability function is given by

$$p(x)=\begin{cases} \left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{1-x} & , \ x=0,1 \\ 0 & , \ otherwise \end{cases}$$

(i) probability of no head $= P(X=0)=\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{1-0}=(1)\left(\frac{1}{2}\right)=\left(\frac{1}{2}\right)$

(ii) Probability of one head $= P(X=1)=\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{1-1}=\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^0=\left(\frac{1}{2}\right)$

3.2.3 Properties of Bernoulli distribution

- Mean of Bernoulli distribution

By definition mean $= E(X)=\sum_x x\,p(x)$

$$=\sum_{x=0}^{1} x\ p^x\ q^{1-x}$$

$$=0p^0q^{1-0}+1p^1q^{1-1}=0+p=p$$

- Variance of Bernoulli distribution

By definition variance $= E(X^2)-[E(X)]^2$

$$=\sum_x x^2 p(x)-(p)^2$$

$$= \sum_{x=0}^{1} x^2 p^x q^{1-x} - p^2$$

$$= 0p^0q^{1-0} + 1p^1q^{1-1} - p^2$$

$$= 0 + p - p^2 = p(1-p) = pq, \ (\because q = 1-p)$$

- Standard deviation $= \sqrt{pq}$

Example 3.6

For a Bernoulli random variable X, the probability of success is equal to 0.6. Find the mean, variance and standard deviation for this Bernoulli probability distribution.

Solution:

(i) By calculation:

X	$p(x)$	$Xp(x)$	$X^2p(x)$
0	0.4	0	0
1	0.6	0.6	0.60
Total	1	0.6	0.60

Mean $= E(X) = \Sigma x\, p(x) = 0.6$

Variance $= V(X) = E(X^2) - [E(X)]^2 = \Sigma x^2\, p(x) - (0.6)^2 = 0.6 - 0.36 = 0.24$

(ii) By properties:

Given that $p = 0.60$, $q = 1 - p = 1 - 0.6 = 0.40$

Mean of Bernoulli distribution $= p = 0.60$

Variance of Bernoulli distribution $= pq = (0.60)(0.40) = 0.24$

S.D$(X) = \sqrt{0.24} = 0.49$

3.3 Binomial experiment

If a Bernoulli trail is repeated a fixed number of times, say n, then such an experiment is called Binomial experiment. It has the following four properties:

i. Each trial has only two possible outcomes i.e. success and failure.
ii. The probability of success remains constant for all trails.
iii. The successive trails are all independent.
iv. The Bernoulli trail is repeated a fixed number of times, say n.

The variable denoting the number of successes of a Binomial experiment is called Binomial random variable i.e. $X = 0, 1, 2, \ldots, n$.

3.3.1 Binomial probability distribution

The binomial probability distribution is given as:

X	0	1	2	...	n	Total
$p(x)$	$\overset{n}{C}_0\, p^0q^{n-0}$	$\overset{n}{C}_1\, p^1q^{n-1}$	$\overset{n}{C}_2\, p^2q^{n-2}$	...	$\overset{n}{C}_n\, p^nq^{n-n}$	1

3.3.2 Binomial probability mass function

The Binomial probability mass function is given by the formula:

$$p(x) = \begin{cases} \overset{n}{C}_x\, p^x q^{n-x}, & x = 0,1,2,\ldots,n \\ 0, & \text{otherwise} \end{cases}$$

This distribution has two parameters n, p whereas n denotes the number of independent Bernoulli trails and p denotes the probability of success on a single Bernoulli trail. Note that if $n = 1$, the Binomial distribution reduces to Bernoulli distribution.

Example 3.7

A fair coin is tossed four times. Find the probability distribution for obtaining various numbers of heads.

Solution:

Here $n = 4$ (four Bernoulli trails)

X: the number of successes in four trials

$X = 0, 1, 2, 3, 4$

$p = \frac{1}{2}$ (probability of head on a single coin)

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

The Binomial probability function is:

$$p(x) = \begin{cases} {}^{4}C_{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x} & , \quad x = 0,1,2,3,4 \\ 0 & , \quad otherwise \end{cases}$$

The Binomial probability distribution is

X	$p(x) = {}^{4}C_{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x}$
0	${}^{4}C_{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{4-0} = \frac{1}{16}$
1	${}^{4}C_{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{4-1} = \frac{4}{16}$
2	${}^{4}C_{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{4-2} = \frac{6}{16}$
3	${}^{4}C_{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{4-3} = \frac{4}{16}$
4	${}^{4}C_{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{4-4} = \frac{1}{16}$
Total	1

Example 3.8

Suppose X has a Binomial distribution with $n = 6$ and $p = 0.75$. Find (i) $P(X = 4)$ (ii) $P(X \geq 5)$ and (iii) $P(X < 4)$.

Solution:

Here $n = 6$

X: number of successes in 6 trials

$X = 0, 1, 2, 3, 4, 5, 6$

$p = 0.75$

$q = 1 - p = -0.75 = 0.25$

The Binomial probability function is

$$p(x) = \begin{cases} {}^{6}C_{x}(0.75)^{x}(0.25)^{6-x} & , \quad x = 0,1,2,3,4,5,6 \\ 0 & , \quad otherwise \end{cases}$$

(i) $P(X = 4) = {}^{6}C_{4}(0.75)^{4}(0.25)^{6-4} = 0.297$

(ii) $P(X \geq 5) = {}^{6}C_{5}(0.75)^{5}(0.25)^{6-5} + {}^{6}C_{6}(0.75)^{6}(0.25)^{6-6} = 0.534$

(iii) $P(X < 4) = {}^{6}C_{0}(0.75)^{0}(0.25)^{6-0} + {}^{6}C_{1}(0.75)^{1}(0.25)^{6-1} + {}^{6}C_{2}(0.75)^{2}(0.25)^{6-2} + {}^{6}C_{3}(0.75)^{3}(0.25)^{6-3} = 0.169$

Example 3.9

A certain drug treatment cures 90% of cases of hookworm in children. Suppose that 20 children suffering from hookworm are to be treated and that the children can be regarded as a sample from the population. Find the probability that (i) all 20 children will be cured (ii) exactly 18 will be cured (iii) at least one will be cured.

Here $n = 20$

X: number of children to be cured

$X = 0, 1, 2, 3 \ldots 20$

$p = 0.90, q = 0.1$

$$p(x) = \begin{cases} \overset{20}{C}_{x}(0.90)^{x}(0.10)^{20-x} & , \quad x = 0,1,2,3,\ldots,20 \\ 0 & , \quad otherwise \end{cases}$$

(i) $P(\text{all 20 will be cured}) = P(X = 20) = \overset{20}{C}_{20}(0.90)^{20}(0.10)^{20-20} = 0.12158$

(ii) $P(\text{exactly 18 will be cured}) = P(X = 18) = \overset{20}{C}_{18}(0.90)^{18}(0.10)^{20-18} = 0.28518$

(iii) $P(\text{at least one will be cured}) = 1 - P(\text{no one will be cured})$

$= 1 - P(X = 0)$

$= 1 - \overset{20}{C}_{0}(0.90)^{0}(0.10)^{20-0}$

$= 1 - 0$ (approximately)

$= 1$

Example 3.10

In Peshawar 42 % of the population has type-A blood. Consider, taking a sample of size 4. Let Y denotes the number of persons in the sample with type-A blood. Find (i) $P(Y = 0)$ (ii) $P(Y = 1)$ (iii) $P(0 \leq Y \leq 2)$ and (iv) $P(0 < Y \leq 2)$

Solution:

Given that $p = 0.42$, $q = 0.58$, $n = 4$

Y: number of persons in the sample with type-A blood

$Y = 0, 1, 2, 3, 4$

$$\therefore p(y) = \begin{cases} \overset{4}{C}_{y}(0.42)^{y}(0.58)^{4-y} & , \quad y = 0,1,2,3,4 \\ 0 & , \quad otherwise \end{cases}$$

(i) $P(Y = 0) = \overset{4}{C}_{0}(0.42)^{0}(0.58)^{4-0} = 0.11316$

(ii) $P(Y = 1) = \overset{4}{C}_{1}(0.42)^{1}(0.58)^{4-1} = 0.32779$

(iii) $P(0 \leq Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2)$

$= 0.11316 + 0.32779 + \overset{4}{C}_{2}(0.42)^{2}(0.58)^{4-2}$

$= 0.44095 + 0.35609$

$= 0.79700$

(iv) $P(0 < Y \leq 2) = P(Y = 1) + P(Y = 2)$

$= 0.32779 + 0.35605 = 0.68384$

Example 3.11

A machine produces 10 per cent defective items. Ten items are selected at random. Find the probability of not more than two items being defective.

Solution:

We have $p = 10\% = \frac{10}{100} = \frac{1}{10}$

$\therefore q = \frac{9}{10}$

$n = 10$

X: number of defective items

$X = 0, 1, 2, 3 \ldots 10$

$$p(x)=\begin{cases}{}^{10}C_x\left(\frac{1}{10}\right)^x\left(\frac{9}{10}\right)^{10-x} & , x=0,1,2,\ldots,10\\ 0 & , \quad otherwise\end{cases}$$

Now

P[not more than two items being defective] $= P[X \le 2]$

$$={}^{10}C_0\left(\frac{1}{10}\right)^0\left(\frac{9}{10}\right)^{10-0}+{}^{10}C_1\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^{10-1}+{}^{10}C_2\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^{10-2}$$

$$=\left(\frac{9}{10}\right)^{10}+10\left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^{9}+45\left(\frac{1}{100}\right)\left(\frac{9}{10}\right)^{8}$$

$$= 0.3487 + 0.3874 + 0.1937 = 0.9298$$

Example 3.12

The chances of a bomber hitting the target and missing the target are 3:2. Calculate the probability that the target will be hit at least once in five sorties.

Solution:

Given that $p=\frac{3}{5}, q=\frac{2}{5}, n=5$

X: denotes the number of hits.

$X = 0, 1, 2, 3, 4, 5$

So $$p(x)=\begin{cases}{}^{5}C_x\left(\frac{3}{5}\right)^x\left(\frac{2}{5}\right)^{5-x} & , x=0,1,2,3,4,5\\ 0 & , \quad otherwise\end{cases}$$

Now, P[at least one will hit the target] $= 1 - P$[no one will hit the target]

$$P(X \ge 1) = 1 - P[X = 0]$$

$$= 1-{}^{5}C_0\left(\frac{3}{5}\right)^0\left(\frac{2}{5}\right)^{5-0}$$

$$= 1-\left(\frac{2}{5}\right)^5 = 1 - 0.01024 = 0.98976$$

3.3.3 Properties of Binomial distribution

(i) Mean of Binomial distribution

By definition

Mean $= E(X) = \sum_x x\,p(x)$

$$= \sum_{x=0}^{n} x\,{}^{n}C_x\,p^x q^{n-x}$$

$$= 0\,{}^{n}C_0\,p^0q^{n-0}+1\,{}^{n}C_1\,p^1q^{n-1}+2\,{}^{n}C_2\,p^2q^{n-2}+\ldots+n\,{}^{n}C_n\,p^nq^{n-n}$$

$$= 0 + npq^{n-1} + \ldots + np^n$$

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

$$= np\,[q + p]^{n-1}$$

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

$$= np$$

(ii) Variance of Binomial distribution

By definition

Variance $= V(X) = E(X^2) - [E(X)]^2$ (A)

As $E(X) = np$ (1)

Now $E(X^2) = \Sigma x^2\,p(x)$

$$= \sum_x [x(x-1) + x]\,p(x) \qquad \because x^2 = x(x-1) + x$$

$$= \sum_x x(x-1)\,p(x) + \sum_x x\,p(x)$$

$$= \sum_{x=0}^{n} x(x-1)\overset{n}{C}_x\, p^x q^{n-x} + \sum_{x=0}^{n} x\ \overset{n}{C}_x\, p^x q^{n-x}$$

$$= \left[0(0-1)\overset{n}{C}_0\, p^0 q^{n-0} + 1(1-1)\overset{n}{C}_1\, p^1 q^{n-1} + 2(2-1)\overset{n}{C}_2\, p^2 q^{n-2} + \ldots + n(n-1)\,p^n q^{n-n}\right] + [np]$$

$$= [0 + 0 + n(n-1)\,p^2 q^{n-2} + \ldots + n(n-1)\,p^n] + (np)$$

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

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

$$= n(n-1)\,p^2\,(1)^{n-2} + (np) \qquad \text{As } p+q=1$$

$$= n^2p^2 - np^2 + np \qquad (2)$$

Putting equation (1) and equation (2) in equation (A) we get

$$V(X) = n^2p^2 - np^2 + np = n^2p^2$$

$$= np - np^2$$

$$= np\,(1-p)$$

$$= npq \qquad (\because q = 1-p)$$

(iii) Standard deviation of Binomial distribution

By definition

$$\text{S.D}(X) = \sqrt{V(X)}$$

$$= \sqrt{npq}$$

Example 3.13

Find the mean, variance and standard deviation of the Binomial distribution whose parameters are $n = 20$ and $p = \frac{3}{5}$.

Solution:

We have $n = 20$, $p = \frac{3}{5}$, $q = \frac{2}{5}$, therefore,

$$\text{Mean} = np = 20 \times \frac{3}{5} = 12$$

$$\text{Variance} = npq = 20 \times \frac{3}{5} \times \frac{2}{5} = 4.8$$

$$\text{S.D}(X) = \sqrt{npq} = \sqrt{4.8} = 2.19$$

Example 3.14

If $n = 4$, $p = \frac{1}{3}$, find (i) the complete Binomial probability distribution (ii) mean and variance of this distribution. (iii) calculate the mean and variance using properties and compare the results.

Solution:

(i) Here $p = \frac{1}{3}$, $q = \frac{2}{3}$, $n = 4$, $X = 0, 1, 2, 3, 4$

$$p(x) = \begin{cases} \overset{4}{C}_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{4-x}, & x = 0,1,2,3,4 \\ 0, & \textit{otherwise} \end{cases}$$

X	$p(x) = {}^{4}C_{x}\left(\frac{1}{3}\right)^{x}\left(\frac{2}{3}\right)^{4-x}$	$Xp(x)$	$X^2p(x)$
0	${}^{4}C_{0}\left(\frac{1}{3}\right)^{0}\left(\frac{2}{3}\right)^{4-0} = \frac{16}{81}$	0	0
1	${}^{4}C_{1}\left(\frac{1}{3}\right)^{1}\left(\frac{2}{3}\right)^{4-1} = \frac{32}{81}$	$\frac{32}{81}$	$\frac{32}{81}$
2	${}^{4}C_{2}\left(\frac{1}{3}\right)^{2}\left(\frac{2}{3}\right)^{4-2} = \frac{24}{81}$	$\frac{48}{81}$	$\frac{96}{81}$
3	${}^{4}C_{3}\left(\frac{1}{3}\right)^{3}\left(\frac{2}{3}\right)^{4-3} = \frac{8}{81}$	$\frac{24}{81}$	$\frac{72}{81}$
4	${}^{4}C_{4}\left(\frac{1}{3}\right)^{4}\left(\frac{2}{3}\right)^{4-4} = \frac{1}{81}$	$\frac{4}{81}$	$\frac{16}{81}$
Total	1	$\frac{108}{81}$	$\frac{216}{81}$

(ii) Mean $= E(X) = \Sigma x\, p(x) = \frac{108}{81} = 1.333$

Variance $= E(X^2) - [E(X)]^2$

$= \Sigma x^2 p(x) - (1.333)^2$

$= \frac{216}{81} - 1.778 = 2.667 - 1.778 = 0.889$

S.D$(X) = \sqrt{0.887} = 0.943$

(iii) Mean and variance by properties:

Mean of Binomial distribution $= n\,p = 4 \times \frac{1}{3} = 1.333$

Variance of Binomial distribution $= n\,p\,q = 4 \times \frac{1}{3} \times \frac{2}{3} = 0.889$

S.D$(X) = 0.943$

Note that both by calculation and properties, we have the same results.

Example 3.15

Is it possible to have a Binomial distribution with mean = 5 and S.D = 4?

Solution:

Given that $n\,p = 5$

Squaring both sides

$$\sqrt{npq} = 4$$

Or $$n\,p\,q = 16$$

$$5q = 16$$

$$q = \frac{16}{5} = 3.2 > 1$$

p or q should not be greater than one because $p + q = 1$

Hence it is not possible to have a Binomial distribution with mean 5 and S.D 4.

Example 3.16

A random variable X is binomially distributed with mean 38 and S.D 2.94. Find n and p.

Solution:

Given $n\,p = 38$ (i)

$$\sqrt{npq} = 2.94$$

Or $npq = 8.64$ (ii)

Putting value of $n\,p$ in equation (ii), we get

$$38q = 8.64$$

$$q = \frac{8.64}{38} = 0.23$$

$$\therefore p = 1 - q = 1 - 0.23 = 0.77$$

Put in equation (i), the value of p, to have

$$n\,(0.77) = 38$$

$$n = \frac{38}{0.77} = 49.351 \cong 50$$

Hence the required parameters are $[n = 49, p = 0.77]$

3.3.4 Binomial frequency distribution

If the Binomial probability distribution is multiplied by N, the number of Binomial experiments, i.e. $Np(x) = N\left[{}^{n}_{x}C\, p^x q^{n-x}\right]$, the resulting distribution is known as Binomial frequency distribution. It is used to find the expected frequency (E_f) of X successes in N Binomial experiments.

Example 3.17

Suppose five fair dice are rolled 96 times. Find the expected frequencies when the number 4, 5 or 6 is regarded as success.

Solution:

We have $N = 96$, $n = 5$, X: number of dice showing 4, 5 or 6.

$X = 0, 1, 2, 3, 4, 5$

$p = P$[resulting 4,5 or 6 on a single die) $= \frac{3}{6} = \frac{1}{2}$, $q = \frac{1}{2}$

$$p(x) = {}^{5}_{x}C\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x},\ x = 0,1,2,3,4,5$$

Now expected frequencies are computed as follows:

X	$p(x) = {}^{5}_{x}C\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x}$	$E_f = N\,p(x)$
0	${}^{5}_{0}C\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{5-0} = \frac{1}{32}$	$96\left(\frac{1}{32}\right) = 03$
1	${}^{5}_{1}C\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{5-1} = \frac{5}{32}$	$96\left(\frac{5}{32}\right) = 15$
2	${}^{5}_{2}C\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{5-2} = \frac{10}{32}$	$= 30$
3	${}^{5}_{3}C\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{5-3} = \frac{10}{32}$	$= 30$
4	${}^{5}_{4}C\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{5-4} = \frac{5}{32}$	$= 15$
5	${}^{5}_{5}C\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^{5-5} = \frac{1}{32}$	$= 03$
Total	1	96

Example 3.18

Fit a Binomial distribution to the following data and compute expected frequencies:

X	0	1	2	3	4
f	30	62	46	10	2

Solution:

First we compute mean of the given frequency distribution

X	0	1	2	3	4	Total
f	30	62	46	10	2	150
fx	0	62	92	30	8	192

Given that $n = 4$, $N = 150$. Now for Binomial distribution $\mu = np$ but here μ is unknown so we replace μ with its estimator $\overline{X}$.

$$\overline{X} = \frac{\sum fx}{\sum f} = \frac{192}{150} = 1.28$$

$$\therefore \overline{X} = np \quad \Rightarrow 1.28 = np$$

$$\text{Or } 1.28 = 4p \quad \Rightarrow p = \frac{1.28}{4} = 0.32 \quad \therefore q = 0.68$$

The fitted Binomial distribution is:

$$p(x) = \begin{cases} \overset{4}{C}_x (0.32)^x (0.68)^{4-x}, & x = 0,1,2,3,4 \\ 0, & \text{otherwise} \end{cases}$$

The expected frequencies are computed as:

X	$p(x) = \overset{4}{C}_x (0.32)^x (0.68)^{4-x}$	$E_f = N\,p(x)$
0	$\overset{4}{C}_0 (0.32)^0 (0.68)^{4-0} = 0.213814$	$150 \times 0.213814 = 32$
1	$\overset{4}{C}_1 (0.32)^1 (0.68)^{4-1} = 0.402478$	$150 \times 0.402478 = 60$
2	$\overset{4}{C}_2 (0.32)^2 (0.68)^{4-2} = 0.284099$	$= 43$
3	$\overset{4}{C}_3 (0.32)^3 (0.68)^{4-3} = 0.089129$	$= 13$
4	$\overset{4}{C}_4 (0.32)^4 (0.68)^{4-4} = 0.010486$	$= 02$
Total	1	150

Example 3.19

Suppose that seven coins are tossed and the number of heads noted. This experiment is repeated 128 times (i) fit a Binomial distribution under the hypothesis that the coins are unbiased. (ii) compute the theoretical frequencies. (iii) find its mean and standard deviation.

Solution:

(i) We have the following information

$n = 7$, $X = 0, 1, 2, 3, 4, 5, 6, 7$ $N = 128$, $p = \frac{1}{2}$ (coins are unbiased),

$q = \frac{1}{2}$, therefore, the fitted Binomial distribution is

$$p(x)=\begin{cases}{}^{7}C_{x}(1/2)^{x}(1/2)^{7-x}, & x=0,1,...,7\\ 0, & otherwise\end{cases}$$

(ii)

X	$p(x)={}^{7}C_{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{7-x}$	$E_f = N\,p(x)$
0	${}^{7}C_{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{7-0}=\frac{1}{128}$	$128\times\frac{1}{128}=01$
1	${}^{7}C_{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{7-1}=\frac{7}{128}$	$128\times\frac{7}{128}=07$
2	${}^{7}C_{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{7-2}=\frac{21}{128}$	$=21$
3	${}^{7}C_{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{7-3}=\frac{35}{128}$	$=35$
4	${}^{7}C_{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{7-4}=\frac{35}{128}$	$=35$
5	${}^{7}C_{5}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{7-5}=\frac{21}{128}$	$=21$
6	${}^{7}C_{6}\left(\frac{1}{2}\right)^{6}\left(\frac{1}{2}\right)^{7-6}=\frac{7}{128}$	$=07$
7	${}^{7}C_{7}\left(\frac{1}{2}\right)^{7}\left(\frac{1}{2}\right)^{7-7}=\frac{1}{128}$	$=01$
Total	1	128

(iii) Mean of Binomial distribution $= E(X) = np = 7\left(\frac{1}{2}\right) = 3.5$

Standard deviation of Binomial distribution $= \sigma_x = \sqrt{npq} = \sqrt{7\times\frac{1}{2}\times\frac{1}{2}}$

$= \sqrt{\frac{7}{4}} = 1.32$

3.4 Hypergeometric experiment

The Binomial distribution is based on the assumption that the successive trials are independent and the probability of success remains unchanged from trial to trial. These assumptions hold only for sampling with replacement from an infinite population but there are experiments in which the conditions of independence is violated and the probability of success does not remain the same/constant for all trails e.g. if sampling is without replacement from a finite population, the probability will change from trail to trail and the successive trails will be dependent. Such experiments are called hypergeometric experiments having the following four properties:

i. Each trail may have two possible results, success and failure.

ii. The probability of success changes on each trail.

iii. The successive trails are dependent.

iv. The experiment is repeated a fixed number of time, say n.

The random variable X denoting the number of successes in a hypergeometric experiment is called hypergeometric random variable and its probability distribution is called hypergeometric probability distribution.

3.4.1 Hypergeometric probability distribution

If values of a hypergeometric random variable along with their associated probabilities are shown in tabular form, then it is called hypergeometric probability distribution i.e.

X	0	1	2	...	n	Total
$p(x)$	$\frac{\binom{K}{0}\binom{N-K}{n-0}}{\binom{N}{n}}$	$\frac{\binom{K}{1}\binom{N-K}{n-1}}{\binom{N}{n}}$	$\frac{\binom{K}{2}\binom{N-K}{n-2}}{\binom{N}{n}}$	...	$\frac{\binom{K}{n}\binom{N-K}{n-n}}{\binom{N}{n}}$	1

3.4.2 Hypergeometric probability mass function

The probability mass function of hypergeometric random variable X is given as:

$$p(x)=\begin{cases}\dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, & x=0,1,2,\ldots,\min(n,K)\\ 0 & \text{otherwise}\end{cases}$$

This distribution has three parameters N, K, n, whereas

- N denotes total number of items in the population.
- K denotes the number of successes in the population.
- n denotes the number of elements in the sample drawn at random.
- X denotes the number of successes in the sample.

Example 3.20

A committee of size three is selected from 4 men and 2 women. Find the probability distribution for the number of women in the committee.

Solution: we have

$$\boxed{\begin{matrix}4M\\2W\end{matrix}} \longleftarrow K$$

$$N \longrightarrow 6 \text{ persons}$$

As it has not been mentioned that an object selected is returned to the population before next draw so, we consider it without replacement sampling case and hence hypergeometric probability distribution is required.

Now $n = 3$ (size of committee)

K=2 (number of successes)

$X = 0, 1, 2$ (number of women in the committee)

Hence the fitted hypergeometric probability function is:

$$p(x)=\begin{cases}\dfrac{\binom{2}{x}\binom{6-2}{3-x}}{\binom{6}{3}}, & x=0,1,2\\ 0, & \text{otherwise}\end{cases}$$

The hypergeometric probability distribution is:

X	$p(x)=\dfrac{\binom{2}{x}\binom{4}{3-x}}{\binom{6}{3}}$
0	$\dfrac{\binom{2}{0}\binom{4}{3-0}}{\binom{6}{3}}=\dfrac{4}{20}$
1	$\dfrac{\binom{2}{1}\binom{4}{3-1}}{\binom{6}{3}}=\dfrac{12}{20}$
2	$\dfrac{\binom{2}{2}\binom{4}{3-2}}{\binom{6}{3}}=\dfrac{4}{20}$
Total	1

Example 3.21

In an international recitation competition of the Holy Quran, a panel of 11 judges is formed to judge the best recitation. Two recitations R_A and R_B were considered to be the best where the opinion of judges got divided. Six judges were in favour of R_A whereas five in favour of R_B. A random sample of five judges was drawn from the panel. Find the probability that out of five judges three are in favour of recitation R_A.

Solution:

In the given problem

$$\boxed{\begin{matrix}6R_A \\ 5R_B\end{matrix}} \longleftarrow K$$

$$N \longrightarrow 11$$

$n = 5, \quad X = 0, 1, 2, 3, 4, 5 \{\because X=0,1,2,\ldots \min(n,k)\}$

$$p(x) = \begin{cases} \dfrac{\binom{6}{x}\binom{11-6}{5-x}}{\binom{N}{n}}, & x = 0,1,2,3,4,5 \\ 0 & \textit{otherwise} \end{cases}$$

Now

$$P(X=3) = \frac{\binom{6}{3}\binom{5}{5-3}}{\binom{11}{5}} = \frac{200}{231}$$

3.4.3 Properties of hypergeometric probability distribution

♦ Mean of hypergeometric probability distribution by definition is given by

$$\text{Mean} = E(X) = \sum_x x\,p(x) = \sum_{x=0}^{n} x\,\frac{\binom{k}{x}\binom{N-k}{n-k}}{\binom{N}{n}}$$

$$= 0\frac{\binom{k}{0}\binom{N-k}{n-0}}{\binom{N}{n}} + \sum_{x=1}^{n} x\,\frac{\binom{k}{x}\binom{N-k}{n-k}}{\binom{N}{n}}$$

$$= \frac{1}{\binom{N}{n}}\sum_{x=1}^{n} x\,\frac{k!}{(k-x)!x!}\binom{N-k}{n-x}$$

$$= \frac{1}{\binom{N}{n}}\sum_{x=1}^{n} x\,\frac{k(k-1)!}{(k-x)!x(x-1)!}\binom{N-k}{n-x}$$

$$= \frac{k}{\binom{N}{n}}\sum_{x=1}^{n}\frac{(k-1)!}{(k-x)!(x-1)!}\binom{N-k}{n-x} = \frac{k}{\binom{N}{n}}\sum_{x=1}^{n}\binom{k-1}{x-1}\binom{N-k}{n-x}$$

Let $y = x-1 \Rightarrow x = y+1$

If $x = 1, y = 0$

If $x = n, y = n-1$

$$\therefore E(X) = \frac{k}{\binom{N}{n}}\sum_{y=0}^{n-1}\binom{k-1}{y}\binom{N-k}{n-y-1}$$

Apply the hypergeometric identity $\sum_{r=0}^{k}\binom{m}{r}\binom{n}{k-r} = \binom{m+n}{k}$

$$E(X) = \frac{k}{\binom{N}{n}}\binom{N-1}{n-1}$$

$$= \frac{K}{\frac{N!}{(N-n)!n!}} \frac{(N-1)!}{(N-n)!(n-1)!} = \frac{Kn(n-1)!}{N(N-1)!}\frac{(N-1)!}{(n-1)!}$$

$$= n\frac{K}{N}$$

♦ Variance of hypergeometric probability distribution

By definition

Variance = $V(X) = E(X^2) - [E(X)]^2$ (A)

As $E(X) = \frac{nK}{N}$.....(1)

Now $E(X^2) = \sum_x x^2 p(x)$

$$= \sum_{x=0}^{n} x^2 \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = (0)^2 \frac{\binom{k}{0}\binom{N-k}{n-0}}{\binom{N}{n}} + \sum_{x=1}^{n} x^2 \frac{\binom{k}{x}\binom{N-k}{N-x}}{\binom{N}{n}}$$

$$= \sum_{x=1}^{n} \left[x(x-1)+x\right]\binom{k}{x}\binom{N-k}{n-x} \quad \because x^2 = x(x-1)+x$$

$$= \sum_{x=1}^{n} x(x-1)\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} + \sum_{x=1}^{n} x\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = \left[1(1-1)\frac{\binom{k}{1}\binom{N-k}{n-0}}{\binom{N}{n}}\right] + \sum_{x=2}^{n} x(x-1)\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} + \frac{nk}{N}$$

$$E(X^2) = \left[\frac{1}{\binom{N}{n}}\sum_{x=2}^{n} x(x-1)\frac{k!}{(k-x)!x!}\binom{N-k}{n-x}\right] + \left(\frac{nk}{N}\right)$$

$$= \frac{1}{\binom{N}{x}}\sum_{x=2}^{n} x(x-1)\frac{k(k-1)(k-2)!}{(k-x)!\,x(x-1)(x-2)!}\binom{N-k}{n-x} + \frac{nk}{N}$$

$$= \frac{k(k-1)}{\binom{N}{n}}\sum_{x=2}^{n}\frac{(k-2)!}{(k-x)!(x-2)!}\binom{N-k}{n-x} + \frac{nk}{N}$$

$$= \frac{k(k-1)}{\binom{N}{n}}\sum_{x=2}^{n}\binom{k-2}{x-2}\binom{N-k}{n-x} + \frac{nk}{N}$$

Let $y = x-2$, $x = y+2$

If $x = 2$, $y = 0$

If $x = n$, $y = n-2$, therefore,

$$E(X^2) = \frac{k(k-1)}{\binom{N}{n}}\sum_{y=0}^{n-2}\binom{k-2}{y}\binom{N-k}{n-y-2} + \frac{nk}{N}$$

Using the hypergeometric identity $\sum_{r=0}^{k}\binom{m}{r}\binom{n}{k-r} = \binom{m+n}{k}$

$$E(X^2) = \frac{k(k-1)}{\binom{N}{n}}\binom{N-2}{n-2} + \frac{nk}{N}$$

$$= \frac{k(k-1)}{\frac{N!}{(N-n)!n!}}\frac{(N-2)!}{(N-n)!(n-2)!} + \frac{nk}{N}$$

$$=\frac{k(k-1)n(n-1)(n-2)!}{N(N-1)(N-2)!}\frac{(N-2)!}{(n-2)!}+\frac{nk}{N}$$

$$=\frac{n(n-1)k(k-1)}{N(N-1)}+\frac{nk}{N}\ldots\ldots\ldots\ldots(2)$$

Putting equation (1) or equation (2) in equation (A)

$$V(X)=\frac{n(n-1)k(k-1)}{N(N-1)}+\frac{nk}{N}-\frac{n^2k^2}{N^2}$$

$$=\frac{nk}{N}\left[\frac{(n-1)(k-1)}{N-1}+1-\frac{nk}{N}\right]$$

$$V(X)=\frac{nk}{N}\left[\frac{N(n-1)(k-1)+N(N-1)-nk(N-1)}{N(N-1)}\right]$$

$$=\frac{nk}{N}\left[\frac{nNk-nN-Nk+N+N^2-N-nkN+nk}{N(N-1)}\right]$$

$$=\frac{nk}{N}\left[\frac{N^2-Nk+nk-nN}{N(N-1)}\right]$$

$$=\frac{nk}{N}\left[\frac{N(N-k)-n(N-k)}{N(N-1)}\right]$$

$$=\frac{nk}{N}\left[\frac{(N-k)(N-n)}{N(N-1)}\right]$$

$$=n\frac{k}{N}\frac{N-k}{N}\frac{N-n}{N-1}$$

$$=n\,p\,q\,\frac{N-n}{N-1}\qquad\left(\because p=\frac{k}{N}, q=1-p=1-\frac{k}{N}=\frac{N-k}{N}\right)$$

♦ Standard deviation of hypergeometric probability distribution is

$$\text{S.D}(X)=\sqrt{npq\frac{N-n}{N-1}}$$

Example 3.22

A bowl contains 4 green and 7 red balls. A sample of 5 balls is selected from the bowl without replacement. Find the probability distribution for the number of green balls. Compute the mean and variance of this probability distribution and compare them with the theoretical mean and variance.

Solution:

The bowl contains

$$\begin{array}{|c|}\hline 4\ G \\ 7\ R \\ \hline\end{array} \longleftarrow k$$

$$N \longrightarrow 11 \text{ balls}$$

Here $n = 5$

X: number of green balls selected

$X = 0, 1, 2, 3, 4.$ [5 is not possible as there are only 4 green balls.].

$$\therefore p(x)=\begin{cases}\dfrac{\binom{4}{x}\binom{11-4}{5-x}}{\binom{11}{5}}, & x=0,1,2,3,4\\ 0 & \textit{otherwise}\end{cases}$$

Now hypergeometric probability distribution of X is given as

X	$p(x)=\frac{\binom{4}{x}\binom{7}{5-x}}{\binom{11}{5}}$	$Xp(x)$	$X^2 p(x)$
0	$\frac{\binom{4}{0}\binom{7}{5-0}}{\binom{11}{5}}=\frac{21}{462}$	0	0
1	$\frac{\binom{4}{1}\binom{7}{5-1}}{\binom{11}{5}}=\frac{140}{462}$	$\frac{140}{462}$	$\frac{140}{462}$
2	$\frac{\binom{4}{2}\binom{7}{5-2}}{\binom{11}{5}}=\frac{210}{462}$	$\frac{420}{462}$	$\frac{840}{462}$
3	$\frac{\binom{4}{3}\binom{7}{5-3}}{\binom{11}{5}}=\frac{84}{462}$	$\frac{252}{462}$	$\frac{756}{462}$
4	$\frac{\binom{4}{4}\binom{7}{5-4}}{\binom{11}{5}}=\frac{7}{462}$	$\frac{28}{462}$	$\frac{112}{462}$
Total	1	$\frac{840}{462}$	$\frac{1848}{462}$

Mean $= E(X) = \sum_{x=0}^{4} x\,p(x) = \frac{840}{462} = 1.8182$

Variance $= V(X) = E(X^2) - [E(X)]^2$

$$= \sum_{x=0}^{4} x^2\,p(x) - (1.8182)^2$$

$$= \frac{1848}{462} - (1.8182)^2 = 4 - 3.3059 = 0.694$$

- Verification

Mean of hypergeometric distribution $= E(X) = \frac{nk}{N} = \frac{5(4)}{11} = \frac{20}{11} = 1.8182$

Variance of hypergeometric distribution $V(X) = \frac{nk}{N}\,\frac{N-k}{N}\,\frac{N-n}{N-1}$

$$= 1.8182\,\frac{(11-4)}{11}\left(\frac{11-5}{11-1}\right)$$

$$= 0.694$$

Key points

- A probability function defined as

$$p(x)=\begin{cases}\frac{1}{N} & , \quad x=1,2,\cdots,N \\ 0 & , \quad otherwise\end{cases}$$

is called discrete uniform probability function

- A random experiment which has two possible outcomes, classified as success and failure, is called a Bernoulli trial.
- If a Bernoulli trail is repeated a fixed number of times, say n, then such an experiment is called Binomial experiment.
- The Binomial probability mass function is given by the formula:

$$p(x)=\begin{cases}{}^{n}C_{x}\, p^{x}q^{n-x} & , \quad x=0,1,2,...,n \\ 0 & , \quad otherwise\end{cases}$$

  - Mean of Binomial distribution is np
  - Binomial distribution has two parameters i.e. n and p. Its range is from 0 to n
  - Variance of Binomial distribution is npq
  - The hypergeometric probability distribution is

$$p(x)=\begin{cases}\dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} & , \quad x=0,1,2,...,\min(n,K) \\ 0 & otherwise\end{cases}$$

  - Hypergeometric distribution has three parameters and its range is from 0 to min(n,k)

Exercise

3.1 Write T for true and F for false in the following statement.

i) If

X	0	1
$p(x)$	$\frac{1}{2}$	$\frac{1}{2}$

then X is not uniform random variable.

ii. Bernoulli experiment will always give only two results.

iii. Tossing a fair coin a large number of times is a Binomial experiment

iv. Binomial probability distribution has only two parameters i.e. n and p.

v. A Binomial random variable is discrete.

vi. Mean and variance of binomial distribution are equal.

vii. If each digit from 0 to 9 has the same probability i.e. $\frac{1}{10}$, they are called random digits.

viii. The Binomial distribution will be symmetrical if $p=q=\frac{1}{2}$

ix. The hypergeometric random variable cannot assume the negative values.

x. Hypergeometric probability distribution has three parameters i.e. N, K, and n.

3.2 Fill in the blanks.

(i) Bernoulli trail has _____ possible outcomes.

(ii) Bernoulli distribution has _____ parameter.

(iii) Range of Binomial random variable is from 0 to _____.

(iv) Parameters of Binomial distribution are _____.

(v) Mean of Binomial distribution is _____ and variance is _____.

(vi) The number of trails in hypergeometric distribution is _____.

(vii) Bernoulli random variable takes only two values i.e. _____.

(viii) Pairs of random digits are called _____.

(ix) The successive trails in a Binomial experiment are _____.

(x) The probability of success _____ from trail to trail in a hypergeometric experiment.

3.3 Choose the correct answer.

i) The Binominal distribution was introduced by

a) Simon Denis Poisson b) Jacob Bernoulli
c) Abraham De Moirés d) R.A. Fisher

ii) In a Bernoulli trail the probability of success is denoted by

a) q b) $1-p$
c) p d) $1-q$

iii) The mean of the Binomial distribution is

a) $\sqrt{np}$ b) npq
c) np d) $\sqrt{npq}$

iv) If X has a Binomial distribution with $p=\frac{2}{3}$ and $n = 9$, then its mean will be equal to

a) 2 b) 3
c) 5 d) 6

v) Hypergeometric probability distribution has parameters

a) 2 b) 3
c) 4 d) 5

vi) The Binomial distribution is the

a) equal probability distribution
b) individual probability distribution
c) discrete probability distribution
d) continuous probability distribution

vii) If $n = 60$, $p = 0.7$ for a Binomial distribution then its standard deviation is equal to

a) 42 b) 6.48
c) 18 d) 3.55

viii) If a fair coin is tossed once, the value of p will be

a) $\frac{1}{2}$ b) $\frac{1}{3}$
c) $\frac{1}{4}$ d) $\frac{1}{1}$

ix) The number of parameters in a Binomial distribution are

a) one b) two
c) three d) four

x) The range of uniform distribution is equal to;

a) $-\infty$ to 0 b) 0,1,...,N
c) 0 to ∞ d) 1, 2, 3,...,N

3.4 Describe in brief the discrete uniform probability distribution.

3.5 What do you know about random digits, random numbers, random number generator and random number table?

3.6 Explain how you would select a random sample of 10 colleges form a list of 206 colleges by using a random number table.

3.7 Define i) Bernoulli experiment ii) Bernoulli random variable iii) Binomial experiment iv) Binomial random variable.

3.8 Define Bernoulli probability distribution. Find its mean, variance and standard deviation.

3.9 i) What is a Binomial experiment and what are its conditions/properties?

ii) Find the mean and variance of the Binomial distribution.

3.10 Suppose X has a Binomial probability distribution with $p = 0.4$ and $n = 3$ Find i) $P(X = -1)$, ii) $P(X = 2)$, iii) $P(X = 1.5)$, iv) $P(X \leq 2)$, v) $P(X \geq 2)$.

3.11 If $n = 5$ and $p = \frac{3}{8}$, Find the complete Binomial probability distribution.

3.12 A fair coin is tossed six times. What is the probability that i) Less than four heads occur ii) 2 or more heads occur.

3.13 If 40% of a consignment of eggs are bad. Estimate the chance that 5 eggs chosen at random contains i) None, ii) one and iii) at least one bad egg.

3.14 A certain drug causes kidney damage 1% of patients. Suppose the drug is to be tested on 50 patients. Find the probability i) none of the patients will experience kidney damage and ii) one or more of the patients will experience kidney damage.

3.15 If 60% of the students in a large college are day-scholar. Find the probability that in a random sample of 12 students from that college exactly 7 will be day-scholar.

3.16 If the probability of hitting a target is $\frac{1}{5}$ and ten shots are fired independently. What is the probability of the target being hit at least twice?

3.17 If the probability that a person entering a utility store will purchase sugar is 0.90. Compute the probability that exactly one person among the next five entering the utility store will purchase sugar?

3.18 a) Define binomial distribution and explain how it arises in practice?

b) Derive its mean and S.D.

3.19 Let the probability of a defective bolt is 0.10, Find a) the mean and b) the standard deviation far the distribution of defective bolts in a total of 400.

3.20 The mean and standard derivation of a Binomial distribution are 3 and 1.5 respectively. Find the two parameters of the Binomial distribution i.e. n and p.

3.21 In a Binomial distribution the mean and variance were found to be 12.38 and 8.64. Find n and p.

3.22 Is it possible to have a Binomial distribution with mean = 5 and standard deviation = 3?

3.23 a) If X is a Binomial random variable with $n = 10$ and $p = 0.6$ then find $E(3X - 2)$.

b) If a Binomial distribution has mean=3 and variance= 2. Find $P(X \leq 5)$

3.24 Find the mean, variance and S.D for the following binomial probability distribution:

X	0	1	2	4	5	6
$p(x)$	0.01521	0.08649	0.21785	0.30833	0.10973	0.02191

Also compare these results with the mean, variance and S.D of the Binomial distribution for $p = 0.5$

3.25 a) Define Binomial frequency distribution.

b) Four dice are thrown and the number of sixes in each throw is recorded, this is repeated 108 times. Find the expected frequencies of 0, 1,2,3,4 sixes.

3.26 Fit a Binomial distribution to the following table.

X	0	1	2	3	4	5	6	7	8	9	10
f	0	1	3	8	16	28	18	13	9	4	0

Compute the theoretical probabilities and find its mean and variance.

3.27 Fit a binomial distribution of the following and compute the expected/theoretical frequencies:

X	0	1	2	3	4	5	6
f	13	70	137	210	145	56	9

3.28 Define hypergeometric experiment, hypergeometric random variable and hypergeometric probability distribution.

3.29 Find the mean and variance of hypergeometric distribution.

3.30 Construct the hypergeometric probability distribution for the number of black balls among 5 balls drawn at random form a box containing 4 white and 7 black balls. Find the mean and variance of this distribution and compare these with the mean and variance of the hypergeometric probability distribution.

3.31 In a manufacturing company 35 employee use touch screen mobile set and 15 have push button sets. Eight employees are selected randomly without replacement. Find the probability that exactly 5 will be using touch screen mobile.

3.32 Four cards are drawn randomly from a well-shuffled deck of 52 playing cards. Calculate the probability that two will be diamond cards.

Unit - 4

Special Continuous Probability Distributions

After studying this unit, the students will be able to

- Define a continuous uniform probability distribution and continuous uniform probability density distribution.
- Find mean, variance and standard deviation of a continuous uniform probability distribution.
- Solve real life problems using continuous uniform probability distribution.
- Define a Normal probability distribution, Normal probability density function, Normal cumulative distribution function, standard normal random variable, standard Normal distribution, standard Normal probability density function and a standard Normal cumulative distribution function.
- Describe the properties of a Normal probability distribution
- Find the ordinates of the standard normal curve using the table of the ordinates of the standard normal curve.
- Find the probabilities for the standard normal random variable using the table of the standard Normal distribution function.
- Inversely use the standard Normal distribution table to determine the value of (i) standard normal random variable corresponding to a given value of the standard Normal cumulative distribution function, (ii) a normal random variable corresponding to a given value of a Normal cumulative distribution function and (iii) parameters of a normal random variable.
- Describe the Normal distribution as a limit of frequency distribution.
- Solve real life problems using Normal probability distribution.

4.1 Introduction to continuous probability distributions

As you know that it is difficult to locate a specified value on a continuous random variable, that is why point probability in continuous case is always equal to zero and we compute probabilities for interval of values. In this unit we consider only continuous uniform distribution and the Normal distribution.

4.1.1 Continuous uniform or rectangular probability distribution

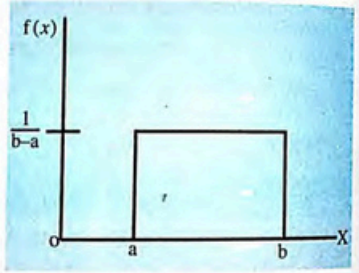
Like discrete case, continuous uniform distribution is the simplest probability distribution. It is used in the situations where the probability density function $f(x)$ remains constant over the entire range of the variable.

4.1.2 Definition of continuous uniform probability density function

The distribution of a continuous random variable X with pdf

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b\\ 0 & otherwise\end{cases}$$

is called continuous uniform probability distribution or probability density function and the variable X is called continuous uniform random variable. This distribution has two parameters, a and b. This distribution is also known as rectangular distribution because its graph is like a rectangle given as.



4.1.3 Properties of continuous uniform probability distribution

(i) Mean of continuous uniform probability distribution

By definition

$$\text{Mean} = E(X)=\int_x x f(x)\,dx$$

$$=\int_a^b x\frac{1}{b-a}dx$$

$$=\frac{1}{b-a}\int_a^b x\,dx$$

$$=\frac{1}{b-a}\left(\frac{x^2}{2}\Big|_a^b\right)$$

$$=\frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right)$$

$$=\frac{1}{b-a}\left\{\frac{(b-a)(b+a)}{2}\right\}$$

$$=\frac{a+b}{2}$$

(ii) Variance of continuous uniform probability distribution

By definition

$$\text{Variance} = V(X) = E(X^2) - [E(X)]^2 \qquad \text{(i)}$$

$$\text{As } E(X) = \frac{a+b}{2} \qquad \text{(ii)}$$

$$E(X^2)=\int_x x^2 f(x)\,dx$$

$$=\int_a^b x^2\frac{1}{b-a}dx$$

$$=\frac{1}{b-a}\int_a^b x^2dx=\frac{1}{b-a}\frac{x^3}{3}\Big|_a^b=\frac{1}{b-a}\left[\frac{b^3-a^3}{3}\right]$$

$$=\frac{1}{b-a}\frac{(b-a)(b^2+ab+a^2)}{3}$$

$$=\frac{b^2+ab+a^2}{3} \qquad \text{(iii)}$$

Put equation (ii) and equation (iii) in equation (i)

$$V(X)=\frac{b^2+ab+a^2}{3}-\frac{(a+b)^2}{4}$$

$$=\frac{4(b^2+ab+a^2)-3(a^2+b^2+2ab)}{12}$$

$$=\frac{4b^2+4ab+4a^2-3a^2-3b^2-6ab}{12}$$

$$=\frac{b^2+a^2-2ab}{12}$$

$$V(X)=\frac{(b-a)^2}{12}$$

(iii) Standard deviation of continuous uniform probability distribution:

By definition

$$S.D(X)=\sqrt{V(X)}=\sqrt{\frac{(b-a)^2}{12}}=\frac{b-a}{\sqrt{12}}$$

Example 4.1

If X has a uniform distribution over the interval (2, 4), find (i) $P(2\le X\le 3)$

(ii) $P(3\le X\le 4)$

Solution:

Given $a = 2$ and $b = 4$, so the pdf of uniform distribution in this case is given by

$$f(x)=\begin{cases}\frac{1}{2} & ,2\le x\le 4\\ 0 & ,otherwise\end{cases}$$

Now

(i) $$P(2\le X\le 3)=\int_2^3\frac{1}{2}dx=\frac{1}{2}\int_2^3 dx$$

$$=\frac{1}{2}\left[x\Big|_2^3\right]=\frac{1}{2}(3-2)=\frac{1}{2}(1)=\frac{1}{2}$$

(ii) $$P(3\le X\le 4)=\int_3^4\frac{1}{2}dx=\frac{1}{2}\int_3^4 dx=\frac{1}{2}\left[x\Big|_3^4\right]$$

$$=\frac{1}{2}[4-3]=\frac{1}{2}(1)=\frac{1}{2}$$

Example 4.2

Let $X\sim U(-1,3)$. Find mean, variance and standard deviation for this continuous uniform random variable.

Solution:

Here $a=-1, b=3$

Now

$$E(X)=\frac{a+b}{2}=\frac{-1+3}{2}=\frac{2}{2}=1$$

$$V(X)=\frac{(b-a)^2}{12}=\frac{[3-(-1)]^2}{12}=\frac{(3+1)^2}{12}=1.333$$

$$S.D(X)=\sqrt{1.333}=1.155$$

4.2 Normal distribution

Normal distribution is the most common and useful amongst all known distribution. It is considered as the cornerstone of the modern statistical theory. The reason of importance is due to the facts that:

i. Many natural phenomena like age, weight, light, I.Q, grade, temperature, income etc. tend to be approximately normal.

ii. Most of the discrete distributions such as Binomial, Poisson, etc.tend to Normal distribution as sample size increases i.e. $n\to\infty$.

iii. For a sample of size $n\ge 30$, the distribution is considered as normal.

iv. Many variables which are not normally distributed can be normalized through suitable transformations.

v. When Normal distribution is shown on a graph paper, it gives a bell-shaped curve called Normal curve which is generally taken as a standard for comparison.

4.2.1 Definition of Normal probability distribution

A random variable X is said to follow a Normal distribution if its probability density function is

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

Where

X = Normal random variable.

$f(x)$ = The height of the curve corresponding to a given value of X.

μ = mean of Normal distribution.

σ = standard deviation of Normal distribution

π = a constant approximately equal to $\frac{22}{7} = 3.1429$

e = a constant approximately equal to 2.7183.

This distribution has two parameters, μ and σ^2. Normal distribution is simply written as $X \sim N(\mu, \sigma^2)$.

4.2.2 Properties of Normal distribution

i) The total probability within its range $-\infty$ to $+\infty$ is always equal to one i.e.

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

ii) The Normal distribution is symmetrical about the mean and is a bell-shaped.

iii) Mean = Median = Mode.

iv) The Normal curve is unimodal.

v) The first and third quartiles are equidistant from the centre, or Mean μ.

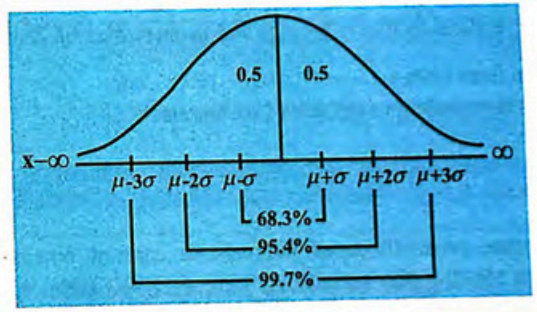
vi) Q.D = $\frac{2}{3}\sigma$ and M.D = $\frac{4}{5}\sigma$

vii) All odd order moments about mean i.e. $\mu_1, \mu_3, \ldots$ are zero.

viii) Measures of skewness, $\beta_1 = 0$ or $\gamma_1 = 0$

ix) Measures of kurtosis, $\beta_2 = 3$ or $\gamma_2 = 0$

x) For Normal distribution, out of the total observation, 68.3 % lies within the limits $(\mu \pm \sigma)$, 95.4 % lies within the limits $(\mu \pm 2\sigma)$ and 99.7 % lies within the limits $(\mu \pm 3\sigma)$. Graphically this statement is shown as



4.2.3 Standard normal variable

Normal random variable X is transformed by subtracting its mean from it and the difference is divided by its standard deviation and is called standard Normal variable, usually denoted by Z, that is

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

Possible values of Z are also form $-\infty$ to $+\infty$.

4.2.4 Definition of standard Normal probability distribution

The probability density function of the standard normal random variable Z defined as

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}, \qquad -\infty \le z \le \infty$$

is called standard Normal probability distribution.

Total probability under the density is equal to one i.e. $\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz=1$. It is simply written as $Z \sim N(0,1)$. The reason for using standard Normal probability distribution in place of Normal probability distribution is that:

(i) To easily calculate the probabilities of Normal distribution i.e. $P[x_1 < X < x_2]=P[z_1 < Z < z_2]$, and probabilities of Z can directly be taken from table 4.1.

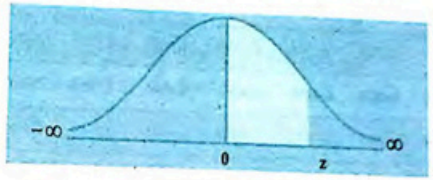
(ii) Z is independent of the unit of measurement.

4.2.5 Area (probability) under the standard normal curve

To compute probabilities of intervals in case of continuous random variables, it is a blessing that statisticians have designed tables in which areas (probabilities) have been compiled. In case of Normal distribution such a table is available for standard Normal probability distribution.

Table 4.1: Areas under the standard Normal curve from 0 to z

The entries in this table are the probabilities that random variable having the standard Normal distribution takes on a value between 0 and z (the shaded area in the figure). For negative values of Z, areas are found by symmetry.



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.0040	0.0080	0.0120	0.0159	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1106	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3990	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4430	0.4441

1.6	0.4452	0.4463	0.4474	0.4484	0.4498	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4719	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.7981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5	0.5	0.5	0.5	0.52	0.5	0.5	0.5	0.5	0.5

4.2.6 Use of the area table

The table of areas of the standard Normal distribution gives probabilities for standard Normal variable Z between 0 (its mean) and a specified value, say z. Therefore, Z-value must always be rounded up to two decimal points as required for reading the table. Locate the first two values in the stub and third value in the box lead. Because of the symmetry property of the Normal distribution the probability (area) between 0 and a positive Z-value must be exactly the same as the probability between a negative Z-value and 0 as long as the Z-value on both sides is of the same magnitude, that is

$$P[0<Z<1.2]=P[-1.2<Z<0]$$

Example 4.3

If Z has a standard Normal distribution, find:

a. $P[-\infty<Z<\infty]$

b. $P[0<Z<\infty]$

c. $P[-\infty<Z<0]$

d. $P[0<Z<2.63]$

e. $P[-1.45<Z<0]$

f. $P[1<Z<1.5]$

g. $P[Z\geq1.64]$

h. $P[-1.96<Z<-1.06]$

i. $P[Z<-1.64]$

j. $P[-1.7<Z<1.25]$

k. $P[Z<-2.46]$

l. $P[Z<2.11]$

Solution:

This example is a practice on how to take the required probability from the areas table. It is quite simple and interesting. First draw standard Normal curve for each case, shade the area in which you are interested and then take the probability directly from table 4.1.

a. $P[-\infty < Z < \infty] = 1$

b. $P[0 < Z < \infty] = 0.5$

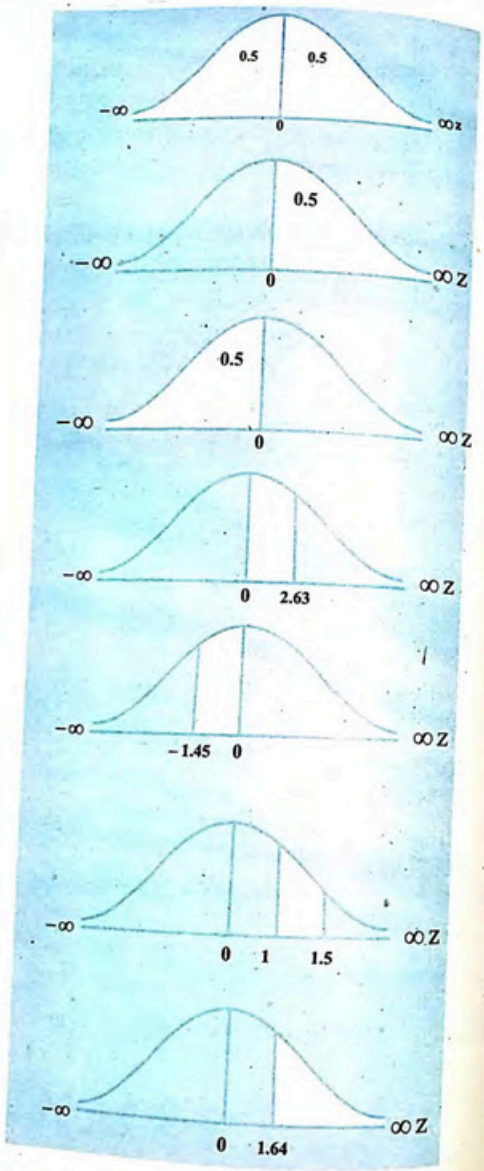
c. $P[-\infty < Z < 0] = 0.5$

d. $P[0 < Z < 2.63] = 0.4957$

e. $P[-1.45 < Z < 0] = 0.4265$

f. $P[1 < Z < 1.5]$
$= P[0 < Z < 1.5] - P[0 < Z < 1]$
$= 0.4332 - 0.3413 = 0.0919$

g. $P[Z \geq 1.64]$
$= P[0 < Z < \infty] - P[0 < Z < 1.64]$
$= 0.5 - 0.4498 = 0.0502$



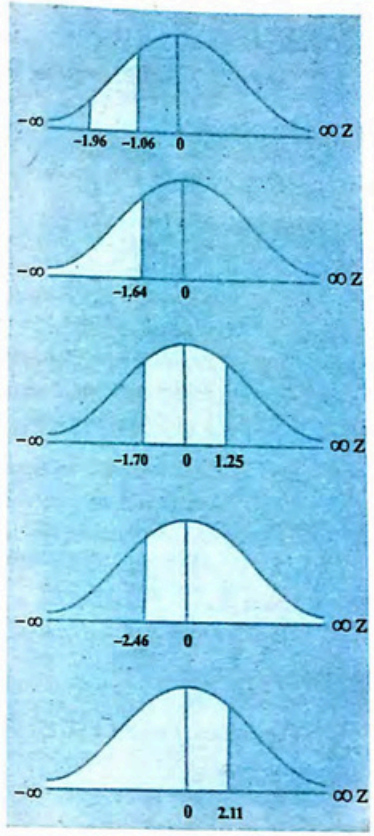
h. $P[-1.96 < Z < -1.06]$
$= P[-1.96 < Z < 0] - P[-1.06 < Z < 0]$
$= 0.4750 - 0.3554 = 0.1196$

i. $P[Z < -1.64]$
$= P[0 < Z < \infty] - P[-1.64 < Z < 0]$
$= 0.5 - 0.4498 = 0.0502$

j. $P[-1.70 < Z < 1.25]$
$= P[-1.70 < Z < 0] + P[0 < Z < 1.25]$
$= 0.4554 + 0.3944 = 0.8498$

k. $P[Z > -2.46]$
$= P[-2.46 < Z < 0] + P[0 < Z < \infty]$
$= 0.4931 + 0.5 = 0.9931$

l. $P[Z < 2.11]$
$= P[-\infty < Z < 0] + P[0 < Z < 2.11]$
$= 0.5 + 0.4826 = 0.9826$

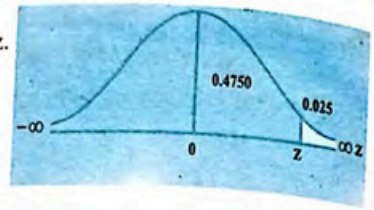


4.2.7 Inverse use of table of areas under standard normal curve

Hope you will be able to determine the interval probability for the standard normal variable Z from the area table. Now we try to reverse the process and determine a value or values of Z corresponding to a given area (probability) from the same areas table. This is called inverse use of the area table.

Example 4.4

If $P[Z \geq z]=0.025$, find the value of z.



Solution:

i) Sketch the given statement graphically as

ii) From the figure we see that the unknown value z is a positive value because it lies to the right of 0 (the mean) and probability above it is equal to 0.025.

iii) As we know that probability from 0 to ∞ is equal to 0.5, so the probability for 0 to z will definitely be equal to 0.5 − 0.025 = 0.4750, i.e. $P[0 \leq Z \leq z]=0.4750$

iv) Now search the figure 0.4750 in the body of the area table to determine the corresponding value of Z. From table 4.1, we observe that the figure 0.4750 corresponds to 1.96. Hence $z = 1.96$ and the area above this value is equal to 0.025.

- Note that if the value we are searching for is not available in the table, then we may pick the nearest value to find Z-values.

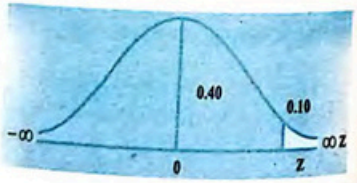
Example 4.5

Determine z when $P[Z \geq z]=0.10$

Solution:

The probability $P[Z \geq z]=0.10$ is graphically shown as:

As $P[Z \geq z] = 0.10 \Rightarrow P[0 < Z < z] = 0.5-0.10 = 0.40$. In the area table 4.1 we observe that exactly 0.40 is not available we consider the closest area 0.3997, which corresponds to a Z-value 1.28. Hence $z = 1.28$ and area above it is equal to 0.10 or 10%.

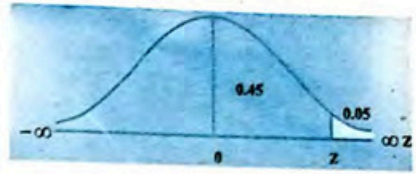


Example 4.6

Let $Z \sim N(0,1)$. If $P[Z \geq z]=0.5$, what will be the value of z?

Solution:

Draw the sketch:



Since $P[Z \geq z] = 0.05$, therefore, $P[0 < Z < z] = 0.5 \quad 0.05 \quad 0.45$. In the area table we can see that 0.45 is lying between 0.4495 and 0.4505 whose corresponding Z-values are 1.64 and 1.65 respectively. The Z-value in this case is given as $z=\frac{1.64+1.65}{2}=1.645$.

Hence $P[Z \geq 1.645]=0.05$

4.2.8 Application of area table to any Normal distribution

After learning the use of areas table and its inverse use, now you would be able to compute probabilities for any Normal random variable X by first converting X to Z by the formula $Z=\frac{X-\mu}{\sigma}$ and then use the table of areas for the standard Normal distribution to obtain the desired probabilities.

Example 4.7

If a Normal distribution has mean 40 and standard deviation 5, find the probabilities for the values of X specified as i) $P(X \geq 44)$, ii) $P(X \leq 25)$, iii) $P(32 < X < 50)$.

Solution:

In order to find the probabilities, we first standardize X by subtracting its mean 40 and dividing the difference by standard deviation 5 to get Z and then find probabilities through area table as follow:

(i) $P[X \geq 44]$

$= P\left[\frac{X-\mu}{\sigma} \geq \frac{44-\mu}{\sigma}\right]$

$= P\left[Z \geq \frac{44-40}{5}\right]$

$= P\left[Z \geq \frac{4}{5}\right]$

$= P[Z \geq 0.8]$

$= P[0 < Z < \infty] - P[0 < Z < 0.80]$

$= 0.5 - 0.2881 = 0.2119$

(ii) $P(X \leq 25)$

$= P\left[\frac{X-\mu}{\sigma} \leq \frac{25-\mu}{\sigma}\right]$

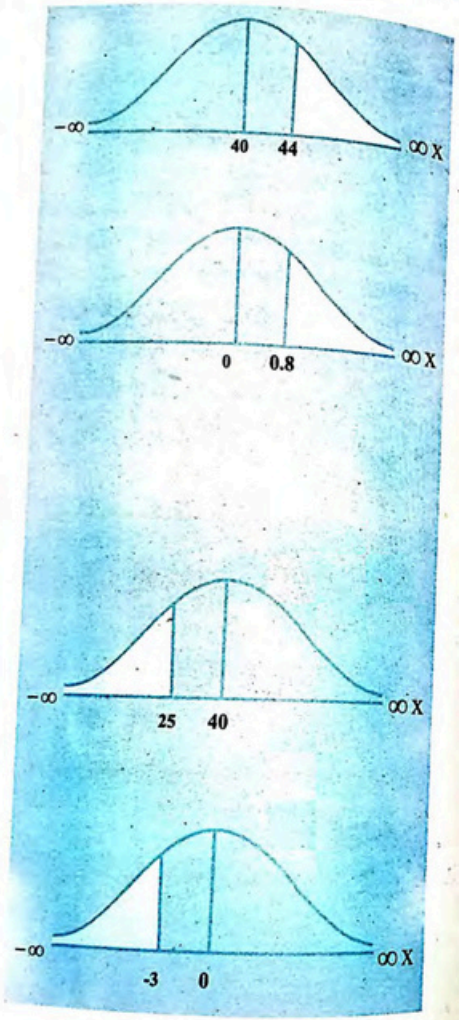
$= P\left[Z \leq \frac{25-40}{5}\right]$

$= P\left[Z \leq \frac{-15}{5}\right]$

$= P[Z < -3]$

$= [-\infty < Z < 0] - P[-3 < Z < 0]$

$= 0.5 - 0.4987 = 0.0013$



(iii) $P(32 \leq X \leq 50)$

$= P\left[\frac{32-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{50-\mu}{\sigma}\right]$

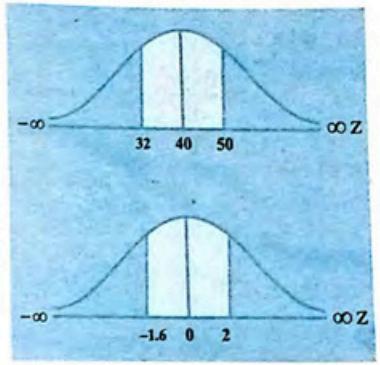
$= P\left[\frac{32-40}{5} < Z < \frac{50-40}{5}\right]$

$= P\left[\frac{-8}{5} < Z < \frac{10}{5}\right]$

$= P[-1.6 < Z < 2]$

$= P[-1.6 < Z < 0] + P(0 < Z < 2)$

$= 0.4452 + 0.4772 = 0.9224$



Example 4.8

Suppose the ages at time of onset of a certain disease are approximately normally distributed with a mean of 11 years and standard deviation of 3 years. A child has just come down with disease. What is the probability that the child is i) between the ages of 8 and 14 years? ii) over 10 years of age? iii) under 12 years?

Solution:

Given $\mu = 11$ years, $\sigma = 3$ years

i) $P[8 < X < 14]$

$= P\left[\frac{8-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{14-\mu}{\sigma}\right]$

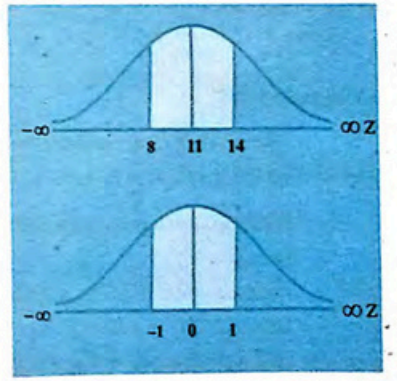
$= P\left[\frac{8-11}{3} < Z < \frac{14-11}{3}\right]$

$= P\left[\frac{-3}{3} < Z < \frac{3}{3}\right]$

$= P[-1 < Z < 1]$

$= P[-1 < Z < 0] + P[0 < Z < 1]$

$= 0.3413 + 0.3413 = 0.6826$



ii) $P[X > 10]$

$= P\left[\frac{X-\mu}{\sigma} > \frac{10-\mu}{\sigma}\right]$

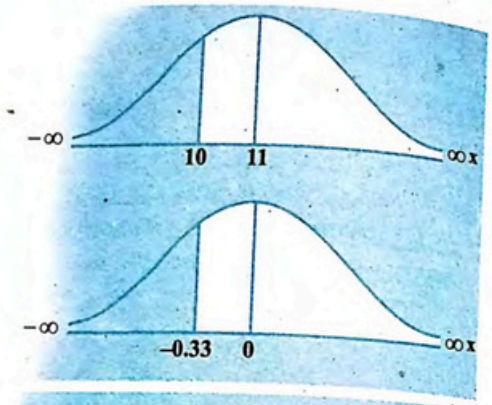
$= P\left[Z > \frac{10-11}{3}\right]$

$= P\left[Z > \frac{-1}{3}\right]$

$= P[Z > -0.33]$

$= P[-0.33 < Z < 0] + P[0 < Z < \infty]$

$= 0.1293 + 0.5 = 0.6293$



iii) $P\ [X < 12]$

$= P\left[\frac{X-\mu}{\sigma} < \frac{12-\mu}{\sigma}\right]$

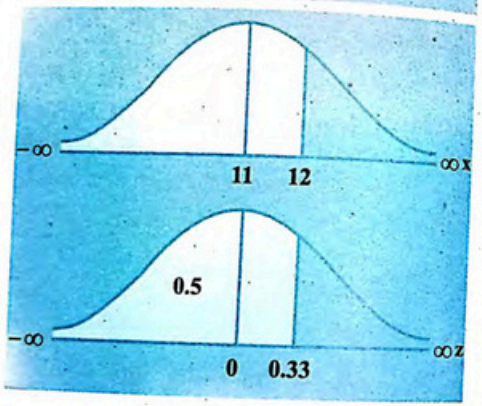
$= P\left[Z < \frac{12-11}{3}\right]$

$= P[Z < 0.33]$

$= P[-\infty < Z < 0] + P[0 < Z < 0.33]$

$= 0.5 + 0.1293$

$= 0.6293$



Example 4.9

The sucrose concentration in a population is normally distributed with a mean = 65 mg and S.D = 25 mg.

i. What proportion of the population is greater than sucrose concentration of 85 mg?

ii. What proportion of the population is less than sucrose concentration of 45 mg?

iii. What proportion of the population are lies between 45 and 85 mg?

Solution:

Given $\mu = 65$ mg

$\sigma = 25$ mg

i) $P[X > 85]$

$= P\left[\frac{X-\mu}{\sigma} > \frac{85-\mu}{\sigma}\right]$

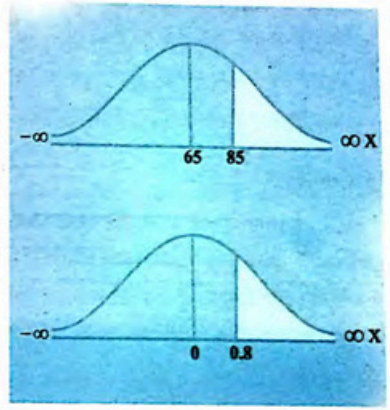
$= P\left[Z > \frac{85-65}{25}\right]$

$= P\left[Z > \frac{20}{25}\right]$

$= P[Z > 0.80]$

$= P[0 < Z < \infty] - P[0 < Z < 0.80]$

$= 0.5 - 0.2881 = 0.2119 = 21.19\%$



ii) $P[X < 45]$

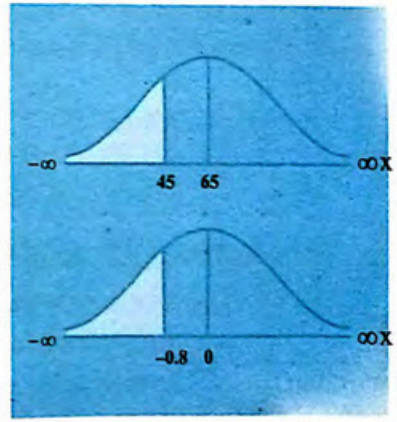
$= P\left[\frac{X-\mu}{\sigma} < \frac{45-\mu}{\sigma}\right]$

$= P\left[Z < \frac{45-65}{25}\right]$

$= P[Z < -0.8]$

$= P[-\infty < Z < 0] - P[-0.80 < Z < 0]$

$= 0.5 - 0.2881 = 0.2119 = 21.19\%$



iii) $P[45 < X < 85]$

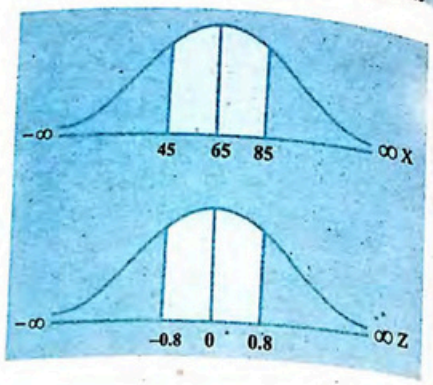
$$= P\left[\frac{45-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{85-\mu}{\sigma}\right]$$

$$= P\left[\frac{45-65}{25} < Z < \frac{85-65}{25}\right]$$

$$= P[-0.80 < Z < 0.80]$$

$$= 2P[0 < Z < 0.80]$$

$$= 2(0.2881) = 0.5762 = 57.62\%$$



Example 4.10

Suppose the yearling trout in a lake have lengths that are approximately normally distributed, about a mean $\mu = 9.5''$ with a standard deviation $\sigma = 1.4''$. (a) What percent of the trout caught have length over 12"? (b) If 80 percent of the trout caught have length greater than x, find x.

Solution:

(a) We need to find

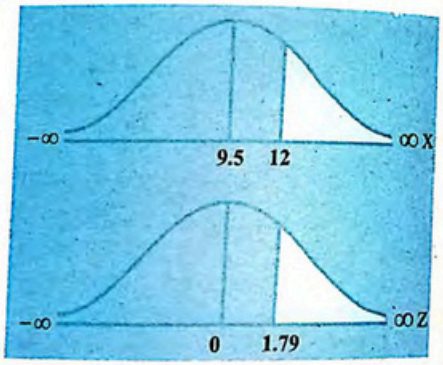
$P[X > 12]$

$$= P\left[\frac{X-\mu}{\sigma} > \frac{12-9.5}{1.4}\right]$$

$$= P\left[Z > \frac{2.5}{1.4}\right] = P[Z > 1.79]$$

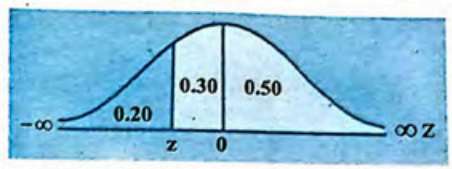
$$= P[0 < Z < \infty] - P[0 < Z < 1.79]$$

$$= 0.5 - 0.4633 = 0.0367 = 3.67\% \quad 4\%$$



(b) We need X-value above which 80% observation will lie.

Mathematically this statement can be written as $P[X > x] = 0.80$ which in standard units is equal to $P[Z > z] = 0.80$, shown in the following figure.



We see that area below z is equal to 0.20. Search this value 0.20 in the area table 4.1, it corresponds $Z = -0.52$ (minus sign is used because Z value lies to the left of 0).

As $Z = \frac{X-\mu}{\sigma}$ $\Rightarrow X = \mu + Z\sigma = 9.5 + (-0.52)\ 1.4 = 9.5 - 0.728 = 8.8$

$X = 8.8$

Hence $P[X > 8.8"] = 0.80 = 80\%$. It means that 80 percent of the trout caught from the lake have length greater than $x = 8.8''$.

Example 4.11

The Peshawar Municipal Corporation installs 10,000 electric lamps in the streets of Peshawar. Average life of these lamps is 1850 hours with a standard deviation of 200 hours. How many lamps may be expected to burn for more than 2000 hours?

Solution:

We first find probability and then multiply by the total number of lamps to compute the required expected frequency as

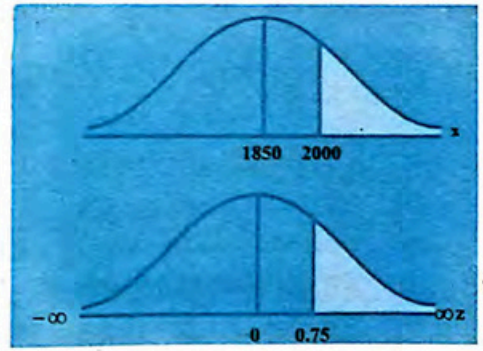
$P[X > 2000]$

$$= P\left[\frac{X-\mu}{\sigma} > \frac{2000-\mu}{\sigma}\right]$$

$$= P\left[Z > \frac{2000-1850}{200}\right]$$

$$= P\left[Z > \frac{150}{200}\right]$$

$$= P[Z > 0.75]$$



$$= P[0 < Z < \infty] - P[0 < Z < 0.75]$$

$$= 0.5 - 0.2734 = 0.2266$$

Thus, the number of lamps that is expected to burn for more than 2000 hours is

$N f(x) = 10{,}000\ (0.2266) = 2266.$

Example 4.12

In a statistics examination, the mean score of students was 78 marks and the standard deviation was 10 marks.

a) Determine the standard scores of students receiving marks:

i) 70 ii) 83 iii) 92.

b) Find the marks corresponding to the standard scores:

i) −1 ii) 1.6.

Solution:

Given $\mu = 78$, $\sigma = 10$

(a) X-values are given and we need to find Z-values:

i) $Z = \frac{X-\mu}{\sigma} = \frac{70-78}{10} = \frac{-8}{10} = -0.8$

ii) $Z = \frac{X-\mu}{\sigma} = \frac{83-78}{10} = \frac{5}{10} = 0.5$

iii) $Z = \frac{X-\mu}{\sigma} = \frac{92-78}{10} = \frac{14}{10} = 1.4$

(b) Z-values are given and X-values are required, we know that

$$Z = \frac{X-\mu}{\sigma}, \Rightarrow X = \mu + z\sigma$$

(i) $X = 78 + (-1)10 = 78 - 10 = 68$

(ii) $X = 78 + (1.6)10 = 78 + 16 = 94$

Example 4.13

Two students A and B were informed that they received standard scores of −1 and 1.6 respectively on a multiple choice examination in Mathematics. If their marks are 68 and 94 respectively, find the mean and standard deviation of the examination marks.

Solution:

We know that, $Z = \frac{X-\mu}{\sigma} \Rightarrow X = \mu + Z\sigma$

Putting values for student A $68 = \mu + (-1)\sigma$

$$68 = \mu - \sigma \quad \text{(i)}$$

student B $94 = \mu + 1.6\sigma$ (ii)

Subtracting equation (i) from equation (ii)

$$26 = 2.6\sigma$$

$$\sigma = \frac{26}{2.6} = 10$$

Putting value of σ in eq (1) we get

$$68 = \mu - 10 \Rightarrow \mu = 78$$

Therefore the mean and standard deviation are $\mu = 78$ and $\sigma = 10$.

4.2.9 Ordinates of the standard Normal distribution

Ordinates mean heights of the standard normal curve corresponding to a specified Z-value. It is denoted by $f(z)$ and is obtained by using the function

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

Fisher and Yates have computed ordinates for different positive values of z and presented them in tabular form as shown in table 4.2. It is important to note that the ordinates at negative values of Z equal to the ordinates at positive values of Z due to unique property of symmetry of the Normal distribution.

TABLE 4.2: Ordinates y or $f(z)$ of the standard Normal curve at z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3812	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.34674	0.3448	0.3429	0.3410	0.3391	0.33742	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.22203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.10401	0.1023	0.1006	0.0989	0.0373	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551

2.0	0.0540	0.0529	0.0519	0.0508	0.0489	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0083	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.00614
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

- **Use of ordinates table**

Ordinates of the standard normal curve given in table 4.2, as usual require Z-value $Z = \frac{X - \mu}{\sigma}$ rounded up to two decimal points.

Example 4.14

Find the ordinates of the standard normal curve at i) $Z = 0.25$ ii) $Z = 3.18$ iii) $Z = -1.64$

Solution:

i) Look at table 4.1 and search 0.2 in the z-column, go in the body of the table across this value up to the column headed by 0.05. The desired ordinate is 0.3867.

ii) Similarly ordinate at $Z = 3.18$ is equal to 0.0025

iii) Ordinate at $Z = -1.64$ = ordinate at $Z = 1.64 = 0.1040$ (By symmetry property).

4.2.10 Fitting of Normal distribution to observed frequency distribution

The fitting of Normal distribution to observe data means to compute theoretical/expected frequencies corresponding to the observed frequencies. This can be done by three methods, the cumulative standard normal probabilities method, the area method and the ordinate method. Here we consider only the ordinate method because it is simple to understand and comparatively less laborious.

♦ Fitting of Normal distribution by ordinate method

This method involves the following steps:

(i) Compute $\bar{x}$ and S from the observed frequency distribution to estimate the unknown parameters μ and σ.

(ii) Calculate $Z = \frac{x - \bar{x}}{S}$ from the mid-points (x).

(iii) Find ordinates corresponding to Z from ordinate table 4.2.

(iv) Multiply the ordinates by $\left(\frac{n\,c}{S}\right)$, to get the expected frequencies, where n is total frequency and c is the size of class interval.

Example 4.15

Fit a Normal distribution to the following table which shows the weight measurements of 60 male workers of a factory.

Weights (kg)	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89
Number of workers	2	4	6	12	16	12	6	2

Solution:

Weights	f	x	fx	fx^2	$z=\frac{x-70.833}{8.029}$	$f(z)$	$e_f=\frac{nc}{s}f(z)=37.37f(z)$
50–54	2	52	104	5408	–2.35	0.0252	37.37(0.0252) = 1
55–59	4	57	228	12996	–1.72	0.0909	3.4
60–64	6	62	372	23064	–1.10	0.2179	8.2
65–69	12	67	804	53868	–0.48	0.3555	13.3
70–74	16	72	1152	82944	0.15	0.3945	15
75–79	12	77	924	71148	0.77	0.2966	11.1
80–84	6	82	492	40344	1.39	0.1518	6
85–89	2	87	174	15138	2.01	0.0529	2
Total	60	–	4250	304910	–	–	60

Here $n = \Sigma f = 60$

$c = 5$

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{4250}{60} = 70.833 \text{ kg}$$

$$S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2} = \sqrt{\frac{304910}{60} - \left(\frac{4250}{60}\right)^2} = \sqrt{5081.83 - 5017.36}$$

$$= \sqrt{64.472} \quad = 8.029$$

As μ and σ are unknown, therefore, $Z = \frac{x - \bar{x}}{S} = \frac{x - 70.833}{8.029}$

Put values of x (mid-points) one by one to get z-values as shown above in the table. Ordinates $f(z)$ is obtained from table 4.2. The factor $\frac{n\,c}{s} = \frac{60\ (5)}{8.029} = 37.37$

Key points

- A density function defined as $f(x) = \begin{cases} \frac{1}{b-a} & , \ a \le x \le b \\ 0 & , \ otherwise \end{cases}$ is the pdf of continuous uniform probability distribution.
- Uniform distribution has two parameters, a and b.
- The range of Normal distribution is from $-\infty$ to $+\infty$.
- For Normal distribution the total probability within its range is always equal to one.
- Normal distribution has two parameters, μ and σ.
- For Normal distribution Mean = Median = Mode.
- For Normal distribution M.D $= \frac{4}{5}\sigma$
- For Normal distribution out of the total observations 68.3 % lies within the limits $\mu \pm \sigma$, 95.4 % lies within the limits $\mu \pm 2\sigma$ and 99.7 % lies within the limits $\mu \pm 3\sigma$.
- $Z = \frac{X - \mu}{\sigma}$ (standard Z-score)
- $P[0 < Z < 1.2] = P[-1.2 < Z < 0]$
- Ordinate mean (height) of the standard normal curve corresponding to a specified z-value. It is denoted by $f(z)$
- Point probability in continuous case is always equal to zero.
- Continuous probability distributions give interval probability.
- When Normal distribution is shown on a graph paper, it gives a bell-shaped curve called normal curve which is generally taken as a standard for comparison.
- For Normal distribution $\beta_1 = 0$ and $\beta_2 = 3$.

Exercise

4.1 Read the following statements carefully and indicate which statement is true or false.

i. If X is a continuous uniform random variable $U(a, b)$ then
$f(x) = \frac{1}{a-b}$, $a < x < b$.

ii. For a Normal distribution, the mean is always lies between the median and the mode.

iii. The right and left tails of the normal curve extend indefinitely, never touching the horizontal axis.

iv. The tail will be on the right hand side of a Normal distribution for any positive z-score.

v. Mean = Median = Mode for Normal distribution.

vi. The mean and the S.D of the Standard Normal distribution is 0 and 1 respectively.

vii. The Normal curve is symmetric around the standard deviation.

viii. For Normal distribution about 99.7% observations lies within the limits $\mu \pm 3\sigma$.

ix. All even order moments of Normal distribution are zero.

x. In case of Normal distribution the two quartiles Q_1 and Q_3 are equidistant from the Centre.

4.2 Fill in the suitable word in the blanks.

i. If X has a continuous random variable over the interval (2, 6), the mean of the distribution is equal to_____

ii. If $X \sim N(\mu, \sigma^2)$, then standard normal variable Z is distributed as _____.

iii. The maximum height of the normal curve lies at the point_____.

iv. For a Normal distribution, the mean deviation from mean is _____.

v. For a Normal distribution, the odd order moments are equal to _____.

vi. The value of skewness = kurtosis = 0 for _____ distribution.

vii. If for a Normal distribution mean = 10, mode = 10, the value of median =?

viii. The normal curve is symmetric around the _____.

ix. The standard normal curve is symmetric around the_____.

x. The interval $(\mu \pm 2\sigma)$ under the normal curve always cover about _____ % of the area.

4.3 Select one correct alternative out of the given ones.

i. The random variable of the Normal distribution is:

(a) discrete (b) continuous (c) positive

ii. If the distribution follows Normal then Mean = Median = Mode

(a) true (b) false (c) impossible

iii. If X is a uniform variable over the interval (5, 10), then the mean of x is

(a) 5 (b) 7.5 (c) 10 (d) 15

iv. Use of the standard normal variable Z instead of normal variable X

(a) complicates the calculation of normal probabilities.

(b) simplifies the calculation of normal probabilities.

(c) Does not make any difference.

(d) gives wrong answer.

v. The total area under a Normal distribution curve to the left of the mean is always equal to

(a) 1 (b) 0 (c) 0.5 (d) 0.9

vi. A normal curve with a small standard deviation will be

(a) positively skewed (b) more spread out

(c) less spread out (d) platy kurtic

vii. The tail of the Normal distribution

(a) meet the horizontal axis at Z = 3.0

(b) cross the horizontal axis at Z = 4.0

(c) never meet or cross the horizontal axis.

(d) are asymmetric

viii. The area under the normal curve within two standard deviation of the mean is

(a) 68.3% (b) 95.4%

(c) 99.7% (d) 99.99%

ix. For a Normal distribution $\mu = 40$, $\sigma = 8$. The value of Z for $X = 52$ is

(a) 2.00 (b) –1.75 (c) 0.80 (d) 1.50

x. An approximate relation between M.D about mean and S.D of Normal distribution is

(a) 5M.D = 4S.D (b) 4M.D = 5S.D

(c) 3M.D = 3S.D (d) 3M.D = 2S.D

4.4 (a) Define continuous uniform probability distribution.

(b) Find mean and variance of the continuous uniform distribution.

4.5 Describe Normal distribution and throw light on the importance of Normal distribution.

4.6 Define Normal distribution. Write down its properties.

4.7 Define standard normal variable and standard Normal distribution. What is the role of this distribution?

4.8 If Z is a standard normal variable, calculate:

(i) $P[Z > 1.60]$

(ii) $P[1.60 < Z < 2.30]$

(iii) $P[-1.64 < Z < -1.02]$

(iv) $P[0 < Z < 1.96]$

(v) $P[-1.96 < Z < 1.96]$

(vi) $P[-1.50 < Z < 0.67]$

(vii) $P[Z < -2.50]$

4.9 Find the proportion of a Normal distribution that corresponds to each of the following sections:

(i) $Z < 0.25$ (ii) $Z > 0.80$ (iii) $Z < -1.50$ (iv) $Z > -0.75$

4.10 Let $Z \sim N(0, 1)$. Find the area under the normal curve in the following cases:

(i) to the right of 2.63 (ii) to the left of –1.45

(iii) between 2.27 and 3.02 (iv) between –1.96 and –1.06

(v) between –2.65 and 2.09 (vi) below 2.17

4.11 Determine the Z-value in the following statements:

(i) $P[Z > z] = 0.005$ (ii) $P[Z > z] = 0.1075$ (iii) $P[Z > z] = 0.9599$

4.12 For a Normal distribution, find the Z-score location that divides the distribution as follows:

(i) separate the top 20% from the rest.

(ii) separate the top 60% from the rest.

(iii) separate the middle 70% from the rest.

4.13 Suppose the distribution of X is normal having mean 46 and standard deviation 4. Compute the following probabilities:

(i) $P[X > 50]$ (ii) $P[X < 38]$ (iii) $P[45 \leq X \leq 49]$

4.14 A Normal distribution has mean = 100 and variance = 225. Find the following probabilities:

(i) $P[X \geq 76]$ (ii) $P[X \geq 124]$ (ii) $P[X \leq 92.5]$

(iv) $P[X \leq 107.5]$ (v) $P[91 \leq X \leq 127]$ (vi) $P[112 \leq X \leq 128.5]$

4.15 The average seasonal rainfall in certain country is 16 inches with a standard deviation of 4 inches. What is the probability that in a year the rainfall in the country will be between 20 inches and 24 inches?

4.16 In a certain book, the frequency distribution of the number of words per page may be taken as approximately normally with mean 400 and SD 25. If a page is chosen at random, what is the probability that the number of words lies between 415 and 450?

4.17 The heights of a certain population of corn plants follow a Normal distribution with mean 145cm and standard deviation 22cm. What percentage of the plant heights are

(i) 100 cm or more, (ii) between 150 and 180 cm (ii) 180cm or more

4.18 The number of calories in a salad on the lunch menu is normally distributed with mean 200 and SD 5. Find the probability that the salad you select will contain

(i) more than 208 calories (ii) between 190 and 200 calories

4.19 A sales tax officer has reported the average sales of the 500 firms that he has to deal with during a year amount to 72,000 with a SD of 20000. Assuming that the sales in these firms are normally distributed, find.

(i) the number of firms whose sales are over 80000 and

(ii) the number of firms whose sales are likely to range between 60000 and 80000.

4.20 The mean height of 1000 students at a certain college is 165 cm and SD is 10cm. Find the number of students whose height is

(i) less than 172 cm (ii) between 159 and 178 cm and

(iii) more than 173.2 cm.

4.21 The mean wage of a certain group of workers working in a factory is Rs 285 with a standard deviation of Rs 50. Find the percentage of workers get above 200 rupees.

4.22 Assume the mean height of soldiers to be 69 inches with a variance of 9 inches. How many soldiers in a regiment of 1000 would you expect to be over six feet tall?

4.23 If $Z \sim N(0,1)$. Find (i) $P(-1 < Z < 1]$, (ii) ordinate of the normal curve for $Z = 2.25$

4.24 A Normal distribution has mean 1 and $\sigma = 3$, find $P[|X| \leq 2]$

4.25 Find the ordinates of the normal curve at

(i) $Z = 2.52$ (ii) $Z = -0.23$ (iii) -1.81

4.26 Find (i) Mean (ii) Standard deviation on an examination in which marks of 90 and 98 correspondents to standard scores of –0.5 and 1.3 respectively.

4.27 Ina Normal distribution the mean is 20 and the standard deviation is 5. What is the approximate value of the mean deviation?

4.28 The mean deviation of a Normal distribution is 16. Find the approximate value of its standard deviation.

4.29 The length of left middle fingers of 1000 criminals is given below:

Length of finger	9.8	10.1	10.4	10.7	11.0	11.3	11.6	11.9	12.2
No. of criminals	4	30	106	206	272	219	120	37	6

Fit a normal distribution to this data by the ordinate method.

4.30 Fit a Normal distribution to the following data.

Classes	40–44	45–49	50–54	55–59	60–64	65–69	70–74
Frequency	9	20	45	55	43	17	11

Unit - 5

Sampling and Sampling Distributions

After studying this unit, the students will be able to

- Define sampling, sampling unit, sampling frame and sample design
- Differentiate between finite and infinite populations, sample survey and complete enumeration. Describe advantages and limitations of sampling.
- Distinguish between probability/random sampling and non-probability/non-random sampling, random sampling with and without replacement.
- Differentiate between sampling and non-sampling errors
- Describe simple random sampling stratified random sampling and systematic random sampling.
- Use the random digits/random numbers table to select a simple random sample from a given finite population.
- Define sampling distribution of statistic and standard error of statistic, sampling distribution of sample mean.
- Describe the properties of a sampling distribution of a sample mean.
- Construct the sampling distribution of sample mean to verify its properties about its mean and variance.
- Define the sampling distribution of difference between two sample means.
- Describe the properties of sampling distribution of difference between two sample means.
- Construct the sampling distribution of difference between two samples means to verify its properties about its mean and variance.
- Define sampling distribution of sample proportion.
- Describe the properties of a sampling distribution of sample proportion.
- Construct the sampling distribution of sample proportion to verify its properties about its mean and variance.
- Define sampling distribution of difference between two sample proportions.
- Describe the properties of sampling distribution of difference between two sample proportions .Construct sampling distribution of difference between two sample proportions and verify its properties about mean and variance.

5.1 Need of sampling survey

The origin of sampling is as old as our civilization and now it is a part of our day-to-day life. Human knowledge and actions are mainly based on samples, specifically, in scientific research. Sample data can either be obtained through sample surveys or experimentation which is the basic input in statistical analysis and inference.

5.1.1 Key terms and definitions

♦ Population

A group of individuals or objects about which we wish to know something is called population or universe.

♦ Finite population

If the number of individuals or items of a population are fixed and limited, it is known as finite population e.g. the population of government hospitals in Pakistan, students in a college, workers in a factory, etc. Finite population usually consists of existing items.

♦ Infinite population

If the population consists of an infinite number of items, it is called infinite population. For example, the population of all real numbers is lying between 10 and 20, the population of stars in the sky etc.

♦ Sampling unit

A single element or group of elements of a population from which required information can be obtained is called sampling unit or unit of the population. For example, if a population has 100 mango trees, then a single mango tree is the sampling unit. A unit may either be natural (e.g. a person, a family, a tree, an animal etc.) or artificial (e.g. a village, districts, a plot of specified size etc.).

♦ Sample

A part or fraction of a population is called sample. In everyday life decision about population is made on the basis of a sample, therefore, it is

important that personal liking or disliking may not be involved during the selection of a sample, that is, sample should be random and must be a true representative of the concerned population.

♦ Survey

To ask a question or a series of questions form many people in order to gather information is called survey. When every individual of the population is examined, it is called population survey or census and if only a part of the population is examined, it is called sample survey or sampling survey. Nowadays sample surveys are being widely used by government departments like agriculture, industries, commerce etc. and private agencies for obtaining estimates of respective parameters. Sample survey has two types i.e. descriptive sample survey (simple information are obtained) and analytical sample survey (comparisons are made between subgroups of the population).

♦ Sampling frame

Before applying any sampling procedure, it is essential to have a list of all the sampling units or a map or other acceptable material which represent the population to be covered. Such a list or map is called sampling frame. For example, if we wish to estimate the wheat crop area in Khyber Pakhtunkhwa, the record of farms along with the names of the farmers, villages etc. are the frame. The frame should not contain inaccurate sampling units and be as up-to-date as possible at the time of use.

♦ Sample design

All principle steps including the methods which are taken in the selection of a sample is called sample design or sample plan or sampling plan. It is formulated before the actual collection of any data. For example, if we want to take a sample of students from Peshawar university, then a complete plan showing number of students to be included in the sample, which students are to be included, the proportion of students, method of sampling to be used, what characteristics are to be studied etc. will be called sample design.

♦ Survey design

The sample design along with some other aspects of the survey e.g. choice and training of interviewers, tabulation plans etc. is called survey design.

5.1.2 Difference between census and sampling

In census (complete enumeration), the information are obtained from each and every unit of the population. In our country census is done after every ten years. Census gives quite reliable information but practically we face the following problems:

i. When population is infinite or area of survey is wide, its study is impossible.
ii. Too much time is required to cover the whole population and often the study becomes out dated by the time it is completed.
iii. Too much resources i.e. money, trained persons etc. are required for survey of the whole population.
iv. When the item or unit is destroyed under investigation, the study of population serves no purpose e.g. testing the life of bulb, battery cell or any other electronic items.

In contrast to census, sampling studies only a selected number of units of the population. For instance (i) A housewife takes one or two grains of rice from the cooking pan and decides whether the rice is cooked or not (ii) A pathologist takes a few drops of blood and tests for any change in blood of the whole body than normal (iii) A quality controller takes a few items and decides whether the lot is in accordance with the desired specifications or not (iv) In a bulb manufacturing factory one tests the life of a few bulbs and then the conclusion about the average life of the whole bulbs produced by the factory is made. All these examples/uses clearly reveal that sampling has been an old practice. Sampling is less expensive and less time consuming.

5.1.3 Advantages of sampling

The most important advantages of sampling over census are;

i. Sampling saves a lot of time because the data are collected and analysed more quickly.

ii. The cost of sampling is very much low than 100 % enumeration because the cost per unit being the same and the number of units in the sample are always less than the number of units in the population.

iii. Sampling results can be more accurate because the sources of error, personal biases and analysis of data can be handled more easily.

iv. When the units are destroyed e.g. testing the life of any electronic instruments, then sampling is the only practical way to assess the average life of the whole lot.

5.1.4 Disadvantages of sampling

Sampling has some limitations given below:

i. When the information is required about every unit of the population, then no sampling method is suitable to give the desired information, only census can do it.

ii. Only large samples can indicate the true characteristics of the population.

iii. Sampling techniques require services of expert persons for better supervision, otherwise, results of the survey may not be reliable.

5.1.5 Parameter, statistic and estimator

Any characteristic of population is called parameter. Parameter is an arbitrary constant value and is usually denoted by Greek letter e.g. μ = mean, σ^2= variance, σ = standard deviation etc. Any function of sample data is called statistic. It is denoted by Latin letter e.g. $\bar{X}$ = mean, S^2 = variance, S = standard deviation etc. A statistic which is used for estimation of parameter is called an estimator. Note that (i) Statistic and estimator are random variables because they vary from sample to sample, (ii) every estimator is a statistic but every statistic is not an estimator.

5.1.6 Sampling and non-sampling errors

♦ Sampling error

The error which occurs due to the natural differences among the members of the population is called sampling error. For example, I.Q, educational level,

income, height, weight, age, etc. of human being are naturally different from each other. Mathematically, this error is written as $e = |\bar{X} - \mu|$ where μ denote parameter and $\bar{X}$ denote estimator. Sampling error can be reduced by (i) increasing the sample size (ii) improving the sample design. It is also called random error or compensating error.

♦ Non-sampling error

The error which occurs during the process of collection and processing the data is called non sampling error. It includes all kinds of human error. Non sampling error can be reduced by (i) employing trained personnel (ii) using modern computational aids.

5.1.7 Bias in sampling

The error which arises due to the personal interest of the investigator is called bias. Error can be minimized in the long run while bias cannot be. This error increases with the increase in sample size. It is also called cumulative error. Some of the factors which introduce bias are (i) deliberate selection of the sample units (ii) substitution of sample units by other units (iii) incomplete or inadequate interviewing (iv) haphazard or accidental selection. Bias cannot be reduced until it is detected.

5.1.8 Sampling with and without replacement

♦ Sampling with replacement

When the item selected for a sample is returned to the population before drawing next item, it is called sampling with replacement (S.W.R) In this case;

i. Possible samples will be N^n

ii. Probability of selection of each unit remains equal.

iii. Sampling units will be independent.

iv. Sampling units can be selected more than once.

v. Population remains infinite.

♦ Sampling without replacement

When the item selected for a sample is not returned to the population before drawing the next item, it is called sampling without replacement (S.W.O.R). In this case;

i. Possible samples will be $\overset{N}{C}_{n}$

ii. Probability of selection of each unit changes from draw to draw.

iii. Sampling units will be dependent.

iv. Sampling units can be selected only once.

v. The sample size is less or equal to the population size, that is, $n \leq N$.

Example 5.1

Suppose a population consist of values 2, 4, 6, 8. Draw all possible samples of size two (i) with replacement and (ii) without replacement.

Solution:

Given 2, 4, 6, 8 (population), $N = 4$ (population size), $n = 2$ (sample size)

(i) Sampling with replacement case:

The number of samples in this case = $N^n = 4^2 = 16$. The procedure for drawing samples is that divide total number of samples by population size i.e. $\frac{16}{4} = 4$ and write every unit of the population 4-times in the first column, then divide the obtained result 4 by the population size i.e. $\frac{4}{4} = 1$ and write every element of the population once in the second column as shown below;

(2, 2)	(4, 2)	(6, 2)	(8, 2)
(2, 4)	(4, 4)	(6, 4)	(8, 4)
(2, 6)	(4, 6)	(6, 6)	(8, 6)
(2, 8)	(4, 8)	(6, 8)	(8, 8)

(ii) Sampling without replacement case

The number of samples in this case = $\overset{N}{C}_{n} = \overset{4}{C}_{2} = 6$. The procedure for drawing samples in this case is that consider the first unit 2 with all other units of the population and keep it aside, then consider 4 with remaining units and keep it aside and so on, we have

(2,4) (2,6) (2,8) (4,6) (4,8) (6,8)

5.1.9 Probability and non-probability sampling methods

Sampling methods are broadly divided in to two categories (i) Probability sampling methods and (ii) non-probability sampling methods.

♦ Probability sampling methods

When each sampling unit of the population has non-zero probability of selection at each draw, the selection procedure is known as probability sampling or random sampling and the sample selected is called random sample. We will describe here the simple random sampling, stratified random sampling and systematic random sampling methods.

♦ Non-probability sampling methods

When selection of units for a sample on each draw is not based on probability but personal judgment plays a significant role, the selection procedure is called non-probability sampling or non-random sampling. Commonly used non-probability sampling methods are: purposive or judgment sampling, Quota sampling, convenience sampling etc.

5.1.10 Simple random sampling

In this method, each and every unit in the population has an equal chance of being selected for a sample. It is also known as random sampling. It is the simplest and widely used method among probability sampling methods. It is a base for other complicated sampling methods. It is the best one for homogenous population. The sample selected by this method is called simple random sample or simply random sample. A random sample may either be selected by with

replacement process with probability $\frac{1}{N^n}$ or without replacement process with probability $\frac{1}{\frac{N}{C_n}}$.

♦ Methods used for selection of a simple random sample

i. Goldfish bowl method

According to this method, serial numbers are allotted from 1 to N to all units of the population. The serial numbers are then written on equal size pieces of paper and are placed in a basket or box or a bowl. After shuffling, a piece of paper is drawn. Its number is noted and is returned to the population in S.W.R case, and then another piece is drawn and so on. The process is continued till the desired number of units is obtained.

ii. Random numbers table method

This method has already been described in unit-3

iii. Computer method

This method is like random number table method. The only difference is that here random numbers are generated by computer packages or calculator.

Example 5.2

Select a simple random sample of size 5, without replacement from the population of 50 employees of a factory by using random number table.

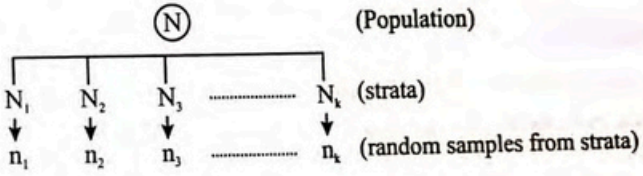
Solution:

Allot two digit serial numbers 00, 01, 02... 49 to employees. Select 5 two digit random number from the random number table, ignoring the two digit number which is greater than 49 or which is appearing again being S.W.O.R case. Let the random numbers be 04, 10, 37, 17, and 48. Thus we will include those employees of the factory in our sample whose serial numbers are 04, 10, 17, 37, and 48.

Note that if we proceed in the same way, we can draw different samples of size 5.

5.1.11 Stratified random sampling

When population is homogenous according to a characteristic e.g. population of first year students is homogenous according to education level, then simple random sampling is good method but when population is heterogeneous with respect to the characteristic in which we are interested e.g. population of students in a postgraduate college is heterogeneous according to the education standard because it contains first year, second year,..., 6th year students. Here simple random sampling does not give satisfactory results and in such situations, stratified random sampling is used to obtain representative sample. According to this method, heterogeneous population is divided in to homogeneous sub-populations, called strata. From each stratum a separate sample is selected by simple random sampling and the overall sample is obtained by combining the samples for all strata. This procedure is known as stratified random sampling and the sample selected by this method is called stratified random sample. For example,



$n_1 + n_2 + \ldots + n_k = n$ and n is called stratified random sample. It is important to note that stratum should be homogenous with respect to the characteristics under study. However, there should be heterogeneity among strata.

♦ Allocation of sample size to various strata

In stratified random sampling one of the major problem is that how much portion of the total sample size n is taken from each stratum? Four methods namely equal allocation, proportional allocation; Neyman's allocation and optimum allocation are generally used for allocation of n to various strata. The

most commonly used one in practice is the proportional allocation method. According to this method, the number of units to be selected from a stratum, is proportional to the size of the stratum and are obtained by the formula:

$$n_h = \frac{n}{N} N_h$$

Where

N = population size

$N_h = h^{th}$ stratum size

n_h = number of units from h^{th} stratum for n.

n = sample size

Example 5.3

Among the 250 employees of a local office 180 are matriculate, 50 are graduates and 20 are master degree holders. If we use proportional allocation to select a stratified random sample of 15 employees, how many employees must we take from each stratum?

Solution:

Given $N = 250, N_1 = 180, N_2 = 50, N_3 = 20, n = 15$

Now random samples from each stratum are obtained by the formula;

$$n_h = \frac{n}{N} N_h$$

$$n_1 = \frac{n}{N} N_1 = \frac{15}{250}(180) = 10.8 \cong 11$$

$$n_2 = \frac{n}{N} N_2 = \frac{15}{250}(50) = 3$$

$$n_3 = \frac{n}{N} N_3 = \frac{15}{250}(20) = 1.2 \cong 1$$

and the stratified random sample is $n = n_1 + n_2 + n_3 = 11 + 3 + 1 = 15$

5.1.12 Systematic random sampling

In systematic random sampling units are selected at equal interval i.e. every k^{th} unit, with a random start. The procedure is that items of the population are first numbered from 1 to N and are then divided into subgroups, each containing $\frac{N}{n} = k$ units, so that $N = nk$. A unit i is selected at random from the first k units, then every k^{th} unit starting with i is selected i.e. i, $(i+k)$, $(i+2k),\ldots,\{i+(n-1)k\}$. Let suppose one is interested in choosing 100 schools from a population of 1000 primary schools. All schools are first given a serial number as 001, 002, 003… 1000. Compute $\frac{N}{n} = \frac{1000}{100} = 10 = k$ .Choose a number at random from the first 10 numbers. Let it is 3, then our sample will include 3rd, 13th, 23rd, 33th…993rd number schools. Systematic random sampling is a type of random sampling. The sample selected by this method is called systematic random sample.

5.2 Sampling distribution

Recall that frequency table or frequency distribution is constructed for summarization of raw data. Probability distribution is constructed for summarizing the values of a random variable. Now, here we are dealing with possible samples and the corresponding estimators like $\bar{X}$, $\hat{p}$, S^2, S etc. The values of these estimators would also need to be summarized in tabular form. This table is technically called sampling distribution. Hence following the same pattern we can define the sampling distribution as; "The probability distribution of all possible values of an estimator is called sampling distribution of that estimator".

5.2.1 Standard error

The standard deviation of the sampling distribution of an estimator is called standard error (S.E). Both have the same concept. The standard error measures the dispersion of all values of an estimator from its average.

5.2.2 Sampling distribution of sample mean

The probability distribution of all possible values of the estimator $\bar{X}$ is called sampling distribution of $\bar{X}$ .i.e.

Sampling distribution of $\bar{X}$

$\bar{X}$	$f(\bar{x})$
$\bar{x}_1$	$f(\bar{x}_1)$
$\bar{x}_2$	$f(\bar{x}_2)$
⋮	⋮
$\bar{x}_k$	$f(\bar{x}_k)$
Total	1

The mean of the sampling distribution of $\bar{X}$ is denoted by $E(\bar{X}) = \mu_{\bar{x}}$ (read as mu sub x bar) and its S.E by $\sigma_{\bar{x}}$ (read as sigma sub x bar).

5.2.3 Properties of the sampling distribution of $\bar{X}$

i. Mean of the sampling distribution of $\bar{X}$ is always equal to the population mean i.e. $\mu_{\bar{x}} = \mu$ (both in with and without replacement sampling)

ii. Standard error of the sampling distribution of $\bar{X}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{(in case of sampling with replacement)}$$

$$= \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{(in case of sampling without replacement)}$$

Note that the factor $\sqrt{\frac{N-n}{N-1}}$ is called finite population correction factor (fpc). It is dropped from the formula when $n < 5\ \%\ N$ and is used when $n \geq 5\ \%\ N$.

iii. Shape of the sampling distribution of $\bar{X}$

There are two cases:

(a) If the sampled population is normal, then shape of the sampling distribution of $\bar{X}$ will also be normal irrespective of the sample size.

(b) If the sampled population is non-normal, then according to central limit theorem (CLT), shape of the sampling distribution of $\bar{X}$ will approximately be normal provided sample size n is large. Note that statisticians consider $n \geq 30$ as large sample.

5.2.4 Formulas for mean and S.E of the sampling distribution of $\bar{X}$

$$\mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) \quad \text{(Mean)}$$

$$\sigma_{\bar{x}}^2 = \sum \bar{x}^2 f(\bar{x}) - \left[\sum \bar{x} f(\bar{x})\right]^2 \quad \text{(Variance)}$$

$$\sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f(\bar{x}) - \left[\sum \bar{x} f(\bar{x})\right]^2} \quad \text{(S.E)}$$

Note that the random variable $\bar{X}$ is standardized as $Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ (In S.W.R.case)

Example 5.4

Consider a family (population) of four children having ages 2, 2, 6, 8 years.

i. Find mean and standard deviation of the population.
ii. Select random samples of two children with replacement and calculate the mean age $\bar{x}$ for each sample.
iii. Construct sampling distribution of $\bar{X}$
iv. Find the mean and standard error of the sampling distribution of mean.
v. Verify the results obtained in (iv) by properties of the sampling distribution of $\bar{X}$

Solution:

(i) Population mean is given by

$$\mu = \frac{\sum x}{N} = \frac{18}{4} = 4.5$$

Population Standard deviation is given by

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2} = \sqrt{\frac{108}{4} - \left(\frac{18}{4}\right)^2} = \sqrt{6.75} = 2.598$$

(ii) Possible samples in S.W.R $= N^n = 4^2 = 16$

S.No.	Samples	$\bar{X}$	S.No.	Samples	$\bar{X}$
1	2, 2	2	9	6, 2	4
2	2, 2	2	10	6, 2	4
3	2, 6	4	11	6, 6	6
4	2, 8	5	12	6, 8	7
5	2, 2	2	13	8, 2	5
6	2, 2	2	14	8, 2	5
7	2, 6	4	15	8, 6	7
8	2, 8	5	16	8, 8	8

(iii) Sampling distribution of $\bar{X}$

$\bar{X}$	Tally bar	f	$f(\bar{x})$
2	IIII	4	4/16
4	IIII	4	4/16
5	IIII	4	4/16
6	I	1	1/16
7	II	2	2/16
8	I	1	1/16
Total		16	1

(iv) Calculation for the mean and S.E of the sampling distribution of $\bar{X}$

$\bar{X}$	$f(\bar{x})$	$\bar{X}f(\bar{x})$	$\bar{X}^2 f(\bar{x})$
2	4/16	8/16	16/16
4	4/16	16/16	64/16
5	4/16	20/16	100/16
6	1/16	6/16	36/16
7	2/16	14/16	98/16
8	1/16	8/16	64/16
Total	1	72/16	378/16

$$\text{Mean} = \mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) = \frac{72}{16} = 4.5$$

$$\text{S.E} = \sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f(\bar{x}) - \left[\sum \bar{x} f(\bar{x})\right]^2}$$

$$= \sqrt{\frac{378}{16} - \left(\frac{72}{16}\right)^2}$$

$$= \sqrt{23.625 - 20.25} = \sqrt{3.375} = 1.84$$

(v) Verification:

(a) As $\mu = 4.5$ and $\mu_{\bar{x}} = 4.5$, hence the property $\mu_{\bar{x}} = \mu$ is satisfied i.e. mean of all sixteen sample means is equal to the population mean.

(b) By property S.E in S.W.R case is

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.598}{\sqrt{2}} = \frac{2.598}{1.414} = 1.84$. It is same as the computed S.E

We learnt that if population mean and standard deviation are known, then it is easy to compute $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ by property. If σ is unknown and n is large then $\sigma_{\bar{x}}$ can be estimated by $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, where s is sample standard deviation.

Example 5.5

A population consists of number 2, 4, 8, 8, 10, 10. Samples of size 2 are to be drawn without replacement from this population.

(i) Find mean and S.D of this population.

(ii) Construct the sampling distribution of $\bar{X}$

(iii) Verify that $E(\bar{X}) = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Solution :

Given population is 2, 4, 8, 8, 10, 10

Here $N = 6$, $n = 2$

(i) $\mu = \frac{\sum x}{N} = \frac{2+4+8+8+10+10}{6} = \frac{42}{6} = 7$

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2} = \sqrt{\frac{348}{6} - \left(\frac{42}{6}\right)^2} = \sqrt{58-49} = \sqrt{9} = 3$$

(iii) Possible samples that could be drawn without replacement $= {}^{6}C_{2} = 15$

S.No.	Samples	$\bar{X}$	S.No.	Samples	$\bar{X}$
1	2, 4	3	9	4, 10	7
2	2, 8	5	10	8, 8	8
3	2, 8	5	11	8, 10	9
4	2, 10	6	12	8, 10	9
5	2, 10	6	13	8, 10	9
6	4, 8	6	14	8, 10	9
7	4, 8	6	15	10, 10	10
8	4, 10	7			

The sampling distribution of $\bar{X}$ is given below:

$\bar{X}$	Tally bar	f	$f(\bar{x})$
3	I	1	1/15
5	II	2	2/15
6	IIII	4	4/15
7	II	2	2/15
8	I	1	1/15
9	IIII	4	4/15
10	I	1	1/15
Total		15	1

(iii) Calculation for mean and S.E of the sampling distribution of $\bar{X}$

$\bar{X}$	$f(\bar{x})$	$\bar{X}f(\bar{x})$	$\bar{X}^2 f(\bar{x})$
3	1/15	3/15	9/19
5	2/15	10/15	50/15
6	4/15	24/15	144/15
7	2/15	14/15	98/15
8	1/15	8/15	64/15
9	4/15	36/15	324/15
10	1/15	10/15	100/15
Total	1	105/15	789/15

$$E(\bar{X}) = \mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) = \frac{105}{15} = 7$$

$$\sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f(\bar{x}) - [\sum \bar{x} f(\bar{x})]^2} = \sqrt{\frac{789}{15} - \left(\frac{105}{15}\right)^2}$$

$$= \sqrt{52.6 - 49} = \sqrt{3.6} = 1.9$$

Now verification of the properties:

(i) As we see that $\mu = 7$, $E(\bar{X}) = 7$, hence proved that $E(\bar{X}) = \mu$

(ii) In S.W.O.R case, the S.E is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-n}} = \frac{3}{\sqrt{2}} \sqrt{\frac{6-2}{6-1}} = \frac{3}{\sqrt{2}} \sqrt{\frac{4}{5}} = 1.9$$

This is exactly the same as computed $\sigma_{\bar{x}}$.

Example 5.6

Draw all possible random samples of size 3 with replacement from a population of three values 3, 6, 9. Obtain the sampling distribution of sample mean and verify the theoretical results.

Solution:

Given population: 3, 6, 9, $N = 3$, $n = 3$

Here sampling is done with replacement, therefore, possible samples are $N^n = 3^3 = 27$. The possible samples and the sample means are listed in the following table.

S.No.	Samples	$\bar{X}$	S.No.	Samples	$\bar{X}$	S.No.	Samples	$\bar{X}$
1	3, 3, 3	3	10	6, 3, 3	4	19	9, 3, 3	5
2	3, 3, 6	4	11	6, 3, 6	5	20	9, 3, 6	6
3	3, 3, 9	5	12	6, 3, 9	6	21	9, 3, 9	7
4	3, 6, 3	4	13	6, 6, 3	5	22	9, 6, 3	6
5	3, 6, 6	5	14	6, 6, 6	6	23	9, 6, 6	7
6	3, 6, 9	6	15	6, 6, 9	7	24	9, 6, 9	8
7	3, 9, 3	5	16	6, 9, 3	6	25	9, 9, 3	7
8	3, 9, 6	6	17	6, 9, 6	7	26	9, 9, 6	8
9	3, 9, 9	7	18	6, 9, 9	8	27	9, 9, 9	9

The sampling distribution

$\bar{X}$	Tally bar	f	$f(\bar{x})$	$\bar{X}f(\bar{x})$	$\bar{X}^2 f(\bar{x})$
3	I	1	1/27	3/27	9/27
4	III	3	3/27	12/27	48/27
5	IIII I	6	6/27	30/27	150/27
6	IIII II	7	7/27	42/27	252/27
7	IIII I	6	6/27	42/27	294/27
8	III	3	3/27	24/27	192/27
9	I	1	1/27	9/27	81/27
Total		27	1	162/27	1026/27

Now $E(\bar{X}) = \mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) = \frac{162}{27} = 6$

And $\sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 f(\bar{x}) - [\sum \bar{x} f(\bar{x})]^2} = \sqrt{\frac{1026}{27} - (6)^2} = \sqrt{38-36} = 1.4142$

Population mean: $\mu = \frac{\sum x}{N} = \frac{3+6+9}{3} = \frac{18}{3} = 6$

Population S.D: $\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2} = \sqrt{\frac{126}{3} - \left(\frac{18}{3}\right)^2} = \sqrt{42-36} = \sqrt{6} = 2.45$

Verification

(i) We see that $E(\bar{X}) = \mu = 6$

(ii) The S.E in S.W.R. case is $\frac{\sigma}{\sqrt{n}} = \frac{2.45}{\sqrt{3}} = 1.4142 = \sigma_{\bar{x}}$

5.3 Sampling distribution of the difference between two sample means

The probability distribution of all possible values of $\bar{X}_1 - \bar{X}_2$ is called sampling distributing of $\bar{X}_1 - \bar{X}_2$. Mean of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is denoted by $\mu_{\bar{x}_1 - \bar{x}_2}$ and S.E by $\sigma_{\bar{x}_1 - \bar{x}_2}$

5.3.1 Properties of the sampling distribution of $\bar{X}_1 - \bar{X}_2$

(i). Mean of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is equal to the difference in population means i.e. $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ (both in S.W.R and S.W.O.R)

(ii). Standard error of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{(in case of S.W.R)}$$

$$= \sqrt{\frac{\sigma_1^2}{n_1}\frac{N_1 - n_1}{n_1 - 1} + \frac{\sigma_2^2}{n_2}\frac{N_2 - n_2}{n_2 - 1}} \quad \text{(in case of S.W.O.R)}$$

(iii). Shape of the sampling distribution of $\bar{X}_1 - \bar{X}_2$

There are two cases:

a) If the sampled population is normal, then sampling distribution of $\bar{X}_1 - \bar{X}_2$ will also be normal irrespective of the samples size.

b) If the sampled population is non-normal, then according to central limit theorem, sampling distribution of $\bar{X}_1 - \bar{X}_2$ will approximately be normal provided n_1 and n_2 both are large.

The variable $\bar{X}_1 - \bar{X}_2$ in standard units is written as

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{\bar{x}_1 - \bar{x}_2})}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \text{(in S.W.R case)}$$

5.3.2 Formulas for $\mu_{\bar{x}_1 - \bar{x}_2}$ and $\sigma_{\bar{x}_1 - \bar{x}_2}$

$$\mu_{\bar{x}_1 - \bar{x}_2} = \sum (\bar{x}_1 - \bar{x}_2) f(\bar{x}_1 - \bar{x}_2)$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sum (\bar{x}_1 - \bar{x}_2)^2 f(\bar{x}_1 - \bar{x}_2) - [\sum (\bar{x}_1 - \bar{x}_2) f(\bar{x}_1 - \bar{x}_2)]^2}$$

Example 5.7

Draw all possible random samples of size $n_1 = 2$ with replacement from a finite population (3, 4, 5) and sample of size $n_2 = 2$ with replacement from another finite population (1, 1, 3).

(i) Find the possible differences between the sample means drawn from the populations.

(ii) Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and standard error.

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

Solution:

Given	Population-1	Population-2
	[3, 4, 5]	[1, 1, 3]
	$N_1 = 3$	$N_2 = 3$
	$n_1 = 2$	$n_2 = 2$

As the case is that of sampling with replacement therefore possible samples from population-1 are $N_1^{n_1} = 3^2 = 9$ and population-2 are $N_2^{n_2} = 3^2 = 9$.

Samples from Population-1		
No	Samples	$\bar{X}_1$
1	3, 3	3
2	3, 4	3.5
3	3, 5	4
4	4, 3	3.5
5	4, 4	4
6	4, 5	4.5
7	5, 3	4
8	5, 4	4.5
9	5, 5	5

Samples from Population-2		
No	Samples	$\bar{X}_2$
1	1, 1	1
2	1, 1	1
3	1, 3	2
4	1, 1	1
5	1, 1	1
6	1, 3	2
7	3, 1	2
8	3, 1	2
9	3, 3	3

(i) The $9 \times 9 = 81$ possible differences $\bar{X}_1 - \bar{X}_2$ are given in the following table

$\bar{X}_2$ \ $\bar{X}_1$	3	3.5	4	3.5	4	4.5	4	4.5	5
1	2	2.5	3	2.5	3	3.5	3	3.5	4
1	2	2.5	3	2.5	3	3.5	3	3.5	4
2	1	1.5	2	1.5	2	2.5	2	2.5	3
1	2	2.5	3	2.5	3	3.5	3	3.5	4
1	2	2.5	3	2.5	3	3.5	3	3.5	4
2	1	1.5	2	1.5	2	2.5	2	2.5	3
2	1	1.5	2	1.5	2	2.5	2	2.5	3
2	1	1.5	2	1.5	2	2.5	2	2.5	3
3	0	0.5	1	0.5	1	1.5	1	1.5	2

(ii) The sampling distribution of $\bar{X}_1 - \bar{X}_2$

$(\bar{X}_1 - \bar{X}_2)$	Tally	f	$f(\bar{x}_1 - \bar{x}_2)$	$(\bar{X}_1 - \bar{X}_2) f(\bar{x}_1 - \bar{x}_2)$	$(\bar{X}_1 - \bar{X}_2)^2 f(\bar{x}_1 - \bar{x}_2)$
0	I	1	1/81	0	0
0.5	II	2	2/81	1/81	0.5/81
1	卌 //	7	7/81	7/81	7/81
1.5	卌 卌	10	10/81	15/81	22.5/81
2	卌 卌 卌 II	17	17/81	34/81	68/81
2.5	卌 卌 卌 I	16	16/81	40/81	100/81
3	卌 卌 卌 I	16	16/81	48/81	144/81
3.5	卌 III	8	8/81	28/81	98/81
4	IIII	4	4/81	16/81	64/81
Total		81	1	189/81	504/81

Now $\mu_{\bar{x}_1-\bar{x}_2} = \sum(\bar{x}_1 - \bar{x}_2) f(\bar{x}_1 - \bar{x}_2) = \frac{189}{81} = 2.33$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\sum(\bar{x}_1 - \bar{x}_2)^2 f(\bar{x}_1 - \bar{x}_2) - \left[\sum(\bar{x}_1 - \bar{x}_2) f(\bar{x}_1 - \bar{x}_2)\right]^2}$$

$$= \sqrt{\frac{504}{81} - \left(\frac{189}{81}\right)^2} = \sqrt{6.222 - 5.429} = \sqrt{0.7931} = 0.88$$

(iii) Verification:

Mean and variance of population 1:

$$\mu_1 = \frac{3+4+5}{3} = \frac{12}{3} = 4$$

$$\sigma_1^2 = \frac{\sum x_1^2}{N_1} - \left(\frac{\sum x_1}{N_1}\right)^2 = \frac{50}{3} - \left(\frac{12}{3}\right)^2 = 16.67 - 16 = 0.67$$

Mean and variance of population 2:

$$\mu_2 = \frac{1+1+3}{3} = \frac{5}{3} = 1.67$$

$$\sigma_2^2 = \frac{\sum x_2^2}{N_2} - \left(\frac{\sum x_2}{N_2}\right)^2 = \frac{11}{3} - \left(\frac{5}{3}\right)^2 = 3.67 - 2.789 = 0.88$$

By property:

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 = 4 - 1.67 = 2.33$$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{0.67}{2} + \frac{0.88}{2}} = \sqrt{0.335 + 0.44} = 0.775 = 0.88$$

These are same as computed results.

5.4 Proportion

An expression which compares a characteristic with the total is called proportion i.e.

$$p = \frac{X}{N} \qquad \text{(parameter)}$$

$$\hat{p} = \frac{X}{n} \qquad \text{(estimator)}$$

Where X denotes the number of individuals having a specified characteristic and is a binomial random variable because each individual has two possibilities i.e. may or may not have the specified characteristic. Thus $X \sim B(n, p)$.

5.4.1 Sampling distribution of sample proportion $\hat{p}$

The probability distribution of all possible values of $\hat{p}$ is called sampling distribution of $\hat{p}$. Mean of the sampling distribution of $\hat{p}$ is denoted by $\mu_{\hat{p}}$ and it's S.E by $\sigma_{\hat{p}}$

5.4.2 Properties of the sampling distribution of $\hat{p}$

(i). Mean of the sampling distribution of sample proportion is always equal to the population proportion i.e. $\mu_{\hat{p}} = p$ (both in S.W.R and S.W.O.R cases)

(ii). S.E of the sampling distribution of $\hat{p}$ is:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \quad , q=1-p \qquad \text{(in case of S.W.R)}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}\frac{N-n}{N-1}} \qquad \text{(in case S.W.O.R)}$$

Note that if p is unknown and n is large, then $\sigma_{\hat{p}}$ can be estimated by $S_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.

(iii). Shape of the sampling distribution of $\hat{p}$

For small n, its shape will be like binomial distribution but for sufficiently large sample sizes, the shape of the sampling distribution of $\hat{p}$ will be approximately normal.

The variable $\hat{p}$ can be standardized as $Z = \frac{\hat{p}-\mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$ (in S.W.R case)

5.4.3 Formulas for $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.

Mean $\quad E(\hat{p}) = \mu_{\hat{p}} = \Sigma \hat{p} f(\hat{p})$

S.E $\quad \sigma_{\hat{p}} = \sqrt{\Sigma \hat{p}^2 f(\hat{p}) - [\Sigma \hat{p} f(\hat{p})]^2}$

Example 5.8

Draw all possible samples of size n = 3 without replacement from the population consists of six numbers 3, 4, 5, 6, 7, 8 and find the proportion of even numbers in the samples. Construct the sampling distribution of sample proportion and verify that;

(a) $\mu_{\hat{p}} = p$

(b) $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}\frac{N-n}{N-1}}$

Solution:

The number of possible samples of size 3 in without replacement case are ${}^{6}C_{3} = 20$. The samples and the proportion of even numbers in the samples are:

S.No.	Samples	$\hat{p}$	S.No.	Samples	$\hat{p}$	S.No.	Samples	$\hat{p}$
1	3, 4, 5	1/3	8	3, 6, 7	1/3	15	4, 6, 8	1
2	3, 4, 6	2/3	9	3, 6, 8	2/3	16	4, 7, 8	2/3
3	3, 4, 7	1/3	10	3, 7, 8	1/3	17	5, 6, 7	1/3
4	3, 4, 8	2/3	11	4, 5, 6	2/3	18	5, 6, 8	2/3
5	3, 5, 6	1/3	12	4, 5, 7	1/3	19	5, 7, 8	1/3
6	3, 5, 7	0	13	4, 5, 8	2/3	20	6, 7, 8	2/3
7	3, 5, 8	1/3	14	4, 6, 7	2/3			

The sampling distribution of $\hat{p}$

$\hat{p}$	f	$f(\hat{p})$	$\hat{p}f(\hat{p})$	$\hat{p}^2 f(\hat{p})$
0	1	1/20	0	0
1/3	9	9/20	3/20	1/20
2/3	9	9/20	6/20	4/20
1	1	1/20	1/20	1/20
Total	20	1	10/20	6/20

$$E(\hat{p}) = \mu_{\hat{p}} = \Sigma \hat{p} f(\hat{p}) = \frac{10}{20} = 0.5$$

$$\sigma_{\hat{p}} = \sqrt{\Sigma \hat{p}^2 f(\hat{p}) - [\Sigma \hat{p} f(\hat{p})]^2}$$

$$= \sqrt{\frac{6}{20} - \left(\frac{10}{20}\right)^2} = \sqrt{0.3 - 0.25} = \sqrt{0.05} = 0.22$$

Verification:

(i) As even numbers in the population are 3 therefore,

$$p = \frac{X}{N} = \frac{3}{6} = \frac{1}{2} = 0.5, \text{ hence proved that } \mu_{\hat{p}} = p$$

(ii) S.E in without replacement case by property is

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}\frac{N-n}{N-1}} = \sqrt{\frac{(0.5)(0.5)}{3}\cdot\frac{6-3}{6-1}} = \sqrt{\frac{0.75}{15}} = \sqrt{0.05} = 0.22$$

This is same as computed value of $\sigma_{\hat{p}}$

5.5 Sampling distribution of difference between two sample proportions

The probability distribution of all possible values of $\hat{p}_1 - \hat{p}_2$ is called sampling distribution of $\hat{p}_1 - \hat{p}_2$. Its mean is denoted by $E(\hat{p}_1 - \hat{p}_2) = \mu_{\hat{p}_1 - \hat{p}_2}$ and S.E by $\sigma_{\hat{p}_1 - \hat{p}_2}$

5.5.1 Properties of the sampling distribution of difference between two sample proportions

(i). Mean of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is equal to the difference between the population proportions i.e.

$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ (both in S.W.R and S.W.O.R)

(ii). S.E of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \quad \text{(S.W.R case)}$$

$$= \sqrt{\frac{p_1 q_1}{n_1}\frac{N_1 - n_1}{N_1 - 1} + \frac{p_2 q_2}{n_2}\frac{N_2 - n_2}{N_2 - 1}} \quad \text{(S.W.O.R case)}$$

If $p_1 \neq p_2$ and also unknown, then for large n_1, n_2 they are replaced with the sample proportions $\hat{p}_1$ and $\hat{p}_2$ respectively. The $\sigma_{\hat{p}_1 - \hat{p}_2}$ is then estimated by

$$S_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

(iii). Shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

For sufficiently large n_1 and n_2 the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

The variable $\hat{p}_1 - \hat{p}_2$ in standard form is written as

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \mu_{\hat{p}_1 - \hat{p}_2}}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \quad \text{(in S.W.R case)}$$

5.5.2 Formulas for mean and S.E of the sampling distribution of $\hat{p}_1 - \hat{p}_2$

$$\mu_{\hat{p}_1 - \hat{p}_2} = \Sigma(\hat{p}_1 - \hat{p}_2) f(\hat{p}_1 - \hat{p}_2)$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\Sigma(\hat{p}_1 - \hat{p}_2)^2 f(\hat{p}_1 - \hat{p}_2) - [\Sigma(\hat{p}_1 - \hat{p}_2) f(\hat{p}_1 - \hat{p}_2)]^2}$$

Example 5.9

Let $\hat{p}_1$ represent the proportion of even numbers in samples of size $n_1 = 2$ drawn with replacement from a finite population consisting units 7, 8, 9. Let $\hat{p}_2$ represent the proportion of odd numbers in samples of size $n_2 = 2$ selected with replacement from another finite population having units 4, 5, and 5. Construct the

sampling distribution of the difference between the sample proportions $\hat{p}_1 - \hat{p}_2$ and verify that $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ and $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

Solution:

Given	population-1	population-2
	[7, 8, 9]	[4, 5, 5]
	$N_1 = 3$	$N_2 = 3$
	$n_1 = 2$	$n_2 = 2$

Since sampling is with replacement, therefore, possible samples are

$N_1^{n_1} = 3^2 = 9$, $N_2^{n_2} = 3^2 = 9$

Population-1		
No	Samples	$\hat{p}_1$
1	(7, 7)	0
2	(7, 8)	1/2
3	(7, 9)	0
4	(8, 7)	1/2
5	(8, 8)	1
6	(8, 9)	1/2
7	(9, 7)	0
8	(9, 8)	1/2
9	(9, 9)	0

Population-2		
No	Samples	$\hat{p}_2$
1	(4, 4)	0
2	(4, 5)	1/2
3	(4, 5)	1/2
4	(5, 4)	1/2
5	(5, 5)	1
6	(5, 5)	1
7	(5, 4)	1/2
8	(5, 5)	1
9	(5, 5)	1

Now $9 \times 9 = 81$ possible differences are given in the following table.

$\hat{p}_1$ \ $\hat{p}_2$	0	1/2	0	1/2	1	1/2	0	1/2	0
0	0	1/2	0	1/2	1	1/2	0	1/2	0
1/2	−1/2	0	−1/2	0	1/2	0	−1/2	0	−1/2
1/2	−1/2	0	−1/2	0	1/2	0	−1/2	0	−1/2
1/2	−1/2	0	−1/2	0	1/2	0	−1/2	0	−1/2
1	−1	−1/2	−1	−1/2	0	−1/2	−1	−1/2	−1
1	−1	−1/2	−1	−1/2	0	−1/2	−1	−1/2	−1
1/2	−1/2	0	−1/2	0	1/2	0	−1/2	0	−1/2
1	−1	−1/2	−1	−1/2	0	−1/2	−1	−1/2	−1
1	−1	−1/2	−1	−1/2	0	−1/2	−1	−1/2	−1

The sampling distribution of $\hat{p}_1 - \hat{p}_2$

$\hat{p}_1 - \hat{p}_2$	f	$f(\hat{p}_1 - \hat{p}_2)$	$(\hat{p}_1 - \hat{p}_2) f(\hat{p}_1 - \hat{p}_2)$	$(\hat{p}_1 - \hat{p}_2)^2 f(\hat{p}_1 - \hat{p}_2)$
-1	16	$\frac{16}{81}$	$-\frac{16}{81}$	$\frac{16}{81}$
$-\frac{1}{2}$	32	$\frac{32}{81}$	$-\frac{16}{81}$	$\frac{8}{81}$
0	24	$\frac{24}{81}$	0	0
$\frac{1}{2}$	08	$\frac{8}{81}$	$-\frac{4}{81}$	$\frac{2}{81}$
1	01	$\frac{1}{81}$	$\frac{1}{81}$	$\frac{1}{81}$
Total	81	1	$\frac{-27}{81} = -\frac{1}{3}$	$\frac{27}{81} = \frac{1}{3}$

Mean of sampling distribution is;

$$\mu_{\hat{p}_1-\hat{p}_2} = \Sigma(\hat{p}_1 - \hat{p}_2) f(\hat{p}_1 - \hat{p}_2) = -\frac{27}{81} = -\frac{1}{3}$$

S.E of sampling distribution is;

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\Sigma(\hat{p}_1 - \hat{p}_2)^2 f(\hat{p}_1 - \hat{p}_2) - \left[\Sigma(\hat{p}_1 - \hat{p}_2) f(\hat{p}_1 - \hat{p}_2)\right]^2}$$

$$= \sqrt{\frac{1}{3} - \left(-\frac{1}{3}\right)^2} = \sqrt{\frac{1}{3} - \frac{1}{9}} = \sqrt{\frac{2}{9}} = 0.47$$

Verification:

There is one even digit in population-1 so $p_1 = \frac{X_1}{N_1} = \frac{1}{3}$

There are two odd digits in population-2 so $p_2 = \frac{X_2}{N_2} = \frac{2}{3}$

Now $p_1 - p_2 = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3} = \mu_{\hat{p}_1-\hat{p}_2}$.

Now by property, the S.E in S.W.R case is

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}{2} + \frac{\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)}{2}} = \sqrt{\frac{1}{9} + \frac{1}{9}} = \sqrt{\frac{2}{9}} = 0.47.$$

Example 5.10

Suppose that 43% of men and 32% of women of a city are in favor of a proposed recreational facility. Describe the sampling distribution of the difference between sample proportions from samples of sizes 350 men and 270 women.

Solution:

Let proportion of men is denoted by p_1 and proportion of women by p_2, then given information is symbolized as;

$p_1 = 43\% = 0.43$ $\quad$ $p_2 = 32\% = 0.32$

$q_1 = 1 - P_1 = 1 - 0.43 = 0.57$ $\quad$ $q_2 = 1 - P_2 = 1 - 0.32 = 0.68$

$n_1 = 350$ $\quad$ $n_2 = 270.$

Describing the sampling distribution of the difference between sample proportions means to find its mean, S.E and the shape.

By property we know that:

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 = 0.43 - 0.32 = 0.11$$

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$= \sqrt{\frac{(0.43)(0.57)}{350} + \frac{(0.32)(0.68)}{270}} = \sqrt{0.0007 + 0.0008} = \sqrt{0.00151} = 0.04$$

Since the samples are sufficiently large, it means that sampling distribution of $\hat{p}_1 - \hat{p}_2$ has approximately normal distribution with mean 0.11 and standard deviation 0.04.

Key points

- A group of individuals or objects about which we wish to know something is called population or universe.
- If the number of individuals or items of a population are fixed and limited, it is known as finite population.
- If the population consists of an infinite number of items it is called infinite population.
- A single element or group of elements of a population from which required information can be obtained is called sampling unit or unit of the population.
- A part or fraction of a population is called sample.
- To ask a question or a series of questions from many people in order to gather information is called survey.
- A list of all the sampling units or a map or other acceptable material which represent the population to be covered. Such a list or map is called sampling frame.
- The sample design along with some other aspects of the survey e.g. choice and training of interviewers, tabulation plans etc. is called survey design.
- Any function of population data is called parameter.
- Any function of sample data is called statistic.
- A statistic which is used for estimation of parameter is called an estimator.
- Every estimator is a statistic but every statistic is not an estimator.
- The error which occurs due to the natural differences among the members of the population is called sampling error.
- The error which occurs during the process of collection and processing the data is called non sampling error.
- The error which arises due to the personal interest of the investigator is called bias.
- The probability distribution of all possible values of an estimator is called sampling distribution of that estimator.
- The standard deviation of the sampling distribution of an estimator is called standard error.
- An expression which compares a characteristic with the total is called proportion

Exercise

5.1 Read the fallowing statements carefully and indicate which statement is true or false.

i. To perform a census, one would need to examine every item in a population under consideration.
ii. A primary objective of sampling is to choose a sample that is representative of the population under consideration.
iii. There is no difference between random error and sampling bias.
iv. An estimator or statistic is characteristic of population.
v. The sample mean and sample standard deviation are not statistics.
vi. As the sample size increases, the S.E decreases.
vii. A sampling procedure that selects items for a sample at uniform intervals is called stratified random sampling.
viii. The probability distribution of all possible means of samples is called sampling distribution of sample mean.
ix. The S.E of the mean is as the S.D of the sampling distribution of the sample mean.
x. S.E can be negative.

5.2 Fill in the blanks.

i. If the number of units in a population is limited, it is known as _____ population.
ii. A population consisting of an unlimited number of units is called a _____ population.
iii. If all units of a population are surveyed, it is called _____.
iv. The discrepancy between a parameter and its estimator due to sampling process is known as _____.
v. Standard deviation of all possible estimates from samples of fixed size is called _____.
vi. The list of all the items of a population is known as _____.

vii. Under simple random sampling with replacement the same unit can occur _____ in the sample.

viii. The expression $\frac{n}{N}$ is known as _____.

ix. The quantity $\sqrt{\frac{N-n}{N-1}}$ is called _____.

x. fpc is dropped from the S.E formula if _____.

5.3 Choose the correct answer.

i. A characteristic of population is called

(a) parameter (b) statistic

(c) sample (d) constant

ii. As the sample size increases the standard error of the sampling distribution of mean:

(a) increases (b) decreases

(c) remains the same (d) becomes negative

iii. Probability of selection varies at each subsequent draw in:

(a) S.W.R (b) S.W.O.R

(c) both (a) and (b) (d) neither (a) and (b)

iv. Which of the following statement is true?

(a) S.E is always zero (b) S.E is always unity

(c) more the S.E, better it is (d) less the S.E, better it is

v. fpc can be dropped from the S.E formula if

(a) $n < 5\% N$ (b) $n < 0.05 N$

(c) $\frac{n}{N} < 5\%$ (d) all of the above

vi. If $\mu = 40, \sigma = 10, n = 25$, then value of $\sigma_{\bar{x}}$ is

(a) 40 (b) 10 (c) 2 (d) 0.4

vii. The number of random samples of size 2 that can be selected W.O.R from a population having 7 items are equal to

(a) 12 (b) 49 (c) 21 (d) 14

viii. A characteristic of sample is called

(a) parameter (b) sample (c) statistic (d) constant

ix. The finite population correction factor is

(a) $\sqrt{\frac{n-1}{n-N}}$ (b) $\sqrt{\frac{N-1}{N}}$ (c) $\sqrt{\frac{N-n}{n-1}}$ (d) $\sqrt{\frac{N-n}{N-1}}$

x. An estimator is a

(a) variable (b) constant

(c) fixed value (d) random variable

5.4 What is meant by population, sample and sampling?

5.5 What is the need of sampling as compared to complete enumeration?

5.6 Write short notes on:

(i) sampling frame (ii) sample design (iii) survey

5.7 Differentiate between:

(i) Finite and infinite population.

(ii) Sampling with and without replacement

(iii) Sampling and non-sampling errors

5.8 Match the symbols on the left with the phrase on the right.

symbol	phrase
μ	Sample mean
$\bar{X}$	Sample variance
σ^2	Population proportion
S^2	Population variance
p	Sample proportion
$\hat{p}$	population mean

5.9 (a) What are the main objectives of sampling?

(b) Define parameter, statistic and estimator.

5.10 Explain sampling and non-sampling errors. What are the methods of reducing these errors?

5.11 Differentiate between probability and non- probability sampling.

5.12 Explain the following methods of selecting a sample:

(i) Simple random sampling

(ii) Stratified random sampling

(iii) Systematic random sampling

5.13 Define a simple random sample. How a random sample could be selected by:

(i) Goldfish bowl method

(ii) Random number table method

5.14 Define the following types of samples:

(i) Random sample (ii) Non-random sample

(iii) Stratified random sample. (iv) Systematic random sample.

5.15 Select a random sample of 10 cities from a list of 200 cities by using a random numbers table.

5.16 (a) Describe stratified random sampling and proportional allocation methods.

(b) The grades in an inter examination of a college were as follows:

Grade	A	B	C	D
No. of students	150	163	195	220

If we wish to select a stratified random sample of size $n = 40$ by proportional allocation, how large a sample must we take from each stratum?

5.17 A stratified random sample of size $n = 200$ is to be taken from a population of size $N = 40{,}000$ divided in to five strata of sizes $N_1 = 15{,}000$, $N_2 = 10{,}000$, $N_3 = 5000$,

$N_4 = 8000$ and $N_5 = 2000$. If the allocation is to be proportional, how large a sample must be taken from each stratum?

5.18 Describe the concept of

(i) Sampling distribution

(ii) Standard error

(iii) Finite population correction factor

5.19 (a) Define the sampling distribution of sample mean and describe its properties.

(b) A population has five elements 4, 5, 6, 7, 8. Draw all possible samples of size 2 with replacement and compute mean for each sample. Construct the sampling distribution of $\bar{X}$ and calculate its mean and S.E also verify that $\mu_{\bar{x}} = \mu$ and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

5.20 A finite population consists of the numbers 2, 2, 4, 6, 5. Select possible random samples of size 2 without replacement and find their means. Construct the sampling distribution of the sample mean and verify that $\mu_{\bar{x}} = \mu$ and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

5.21 A population has a mean of $\mu = 50$ and a standard deviation of $\sigma = 12$. For samples of size $n = 4$, what is the mean (expected value) and the standard deviation (standard error) for the distribution of sample mean?

5.22 For a population with a mean of $\mu = 40$ and a standard deviation of $\sigma = 8$, find the Z-score corresponding to a sample mean of $\bar{x} = 44$ for each of the following sample sizes:

(i) $n = 4$(ii) $n = 16$

5.23 Given the five-element population 4, 5, 7, 9, 10.

(i) Compute population mean and variance.

(ii) Suppose samples of size $n = 3$ are selected without replacement but you compute $\sigma_{\bar{x}}^2$ directly.

(iii) Suppose samples of size $n = 3$ are selected with replacement but you compute $\sigma_{\bar{x}}^2$ directly.

5.24 Draw all possible samples of size $n = 3$ without replacement from the population 0, 3, 6, 12, 15, 18. Construct the sampling distribution of the sample mean and verify the relation between:

(i) Mean of the sampling distribution and the population mean.

(ii) Standard deviation of the sampling distribution of the mean and the population standard deviation.

5.25 Define sampling distribution of difference between means of two samples. Describe its important properties.

5.26 Draw all possible random samples of size $n_1 = 2$ with replacement from the finite population consisting of –2, 0, 2/and 4. Similarly draw all possible random samples of size $n_2 = 2$ with replacement from the population –1 and +1.

(i) Find the possible differences between the sample means of the two populations.

(ii) Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$.

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

5.27 Draw all possible random samples of size $n_1 = 2$ without replacement from a finite population consisting of 3, 6, 9. Similarly draw all possible random samples of size $n_2 = 2$ without replacement from another finite population consisting of 2, 4, 6.

(i) Find the possible differences between the sample means of the two populations.

(ii) Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and variance. Also verify the theoretical results.

5.28 What is meant by proportion and sampling distribution of sample proportion $\hat{p}$? Describe its important properties.

5.29 Draw all possible samples of size $n = 3$ without replacement from the population 2, 3, 3, 4, 5, 6 and find the sample proportion $\hat{p}$ of odd numbers in the samples. Construct the sampling distribution of sample proportion and verify that $\mu_{\hat{p}} = p$ and $Var(\hat{p}) = \frac{pq}{n}\left[\frac{N-n}{N-1}\right]$

5.30 The marital status of a population of seven friends is U, M, M, U, M, U, U where U and M stand for unmarried and married respectively. Find the proportion of married friends in the population. Take all possible samples of two friends without replacement from this population and find the proportion of married friends in each sample. Make the sampling distribution of the sample proportion and verify that $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}}^2 = \frac{pq}{n}\left(\frac{N-n}{N-1}\right)$.

5.31 Let $\hat{p}_1$ represent the proportion of even numbers in a random sample of size $n_1 = 2$ without replacement from a finite population consisting of values 4, 6, 9. Similarly, let $\hat{p}_2$ represent the proportion of even numbers in a random sample of size $n_2 = 2$ without replacement from another finite population consisting of values 2, 2, 5. Form a sampling distribution of $\hat{p}_1 - \hat{p}_2$ and verify that:

(i) $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

(ii) $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1}\frac{N_1 - n_1}{N_1 - 1} + \frac{p_2 q_2}{n_2}\frac{N_2 - n_2}{N_2 - 1}$

Unit - 6

Estimation

After studying this unit, the students will be able to

- Define estimation of a parameter, point estimation of a parameter, point estimator and point estimate.
- Differentiate between point estimator and point estimate.
- Describe from a random sample, the point estimators and point estimate for population mean and variance.
- Define un-biasedness, un-biased estimator, biased estimator and biased.
- Describe the methods to reduce bias in sample surveys.
- Describe and verify the un-biasedness of sample mean, sample proportion and sample variance.
- Use calculator in statistical mode to directly find the un-biased estimates of mean and variance of the population from which the sample was drawn.
- Define efficiency and explain best estimator.
- Identify the best estimator of population mean, population variance and population proportion.
- Find the best estimates of population mean, population variance and population proportion from a given random sample.
- Identify the pooled estimators, from two samples, of population mean, population variance and population proportion.
- Find the pooled estimates of population mean, population variance and population proportion from two given random samples.
- Define interval estimation of a parameter, interval estimate and confidence coefficient.
- Explain and estimate the confidence interval for the mean of a normal population (known and unknown standard deviations), the difference between means of two normal populations (known and unknown standard deviations), the population proportion (large sample) and the difference between proportions of two populations (large samples).

6.1 Introduction to statistical inference

Statistical inference is a process of drawing conclusions (inferences) about the population on the basis of sample information obtained from that population. There are two branches of statistical inference.

i. Estimation of parameters
ii. Hypothesis testing

Statistical inference is of immense importance because complete knowledge regarding population is seldom available. In the previous unit concepts of sampling and sampling distributions were discussed which is actually a base for statistical inference i.e. sampling allow us to make use of the information gathered for the sample to draw inferences about the entire population.

6.1.1 Estimation of parameters

Statistical estimation is a process by which the unknown value of a parameter is obtained from the sample observations. Suppose we want to know the average age of people in our country, the percentage of smokers in the Khyber Pakhtunkhwa etc. then these are the problems which come under estimation.

6.1.2 Types of estimation

There are two types of estimation (i) point estimation (ii) interval estimation. When a specific value is obtained from sample observations and is used to estimate the unknown value of the parameter, the process is called point estimation and the single estimated value is called point estimate or simply estimate. For example, the values obtained by $\bar{X}$, S^2, $\hat{p}$ are point estimates for the parameters μ, σ^2, p respectively. When a range of values is obtained from sample observations within which the unknown value of parameter is believed to lie, the process is called interval estimation and the resulting interval of two numbers is called interval estimate.

6.1.3 Difference between an estimator and an estimate

A rule, usually expressed as a formula that tells us how to calculate an estimate from the sample data is called an estimator and the resulting number is called an estimate.

For example; if $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = 100$ then $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ is an estimator and 100 is an estimate. Similarly, if $s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = 2.63$ then $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ is an estimator and its numerical value 2.63 is an estimate.

6.2 Point estimation

This process provides a single value which is calculated from the sample data as an estimate for the unknown population parameter. Point estimate may or may not be close to the parameter because the random sample used is one of the many possible samples which could be selected from the population.

Example 6.1

A random sample has ten observations 4, 3, 7, 8, 4, 10, 5, 5, 4 and 9. Compute a point estimate of (i) population mean (ii) population standard deviation (iii) standard error of the mean (iv) population proportion of even numbers.

Solution:

(i) Point estimate of the population mean μ is;

$$\bar{X} = \frac{\sum x}{n} \qquad \text{(estimator)}$$

$$= \frac{59}{10} = 5.9 \qquad \text{(point estimate)}$$

(ii) Point estimate of the population standard deviation σ is;

$$s = \sqrt{\frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]} \qquad \text{(estimator)}$$

$$= \sqrt{\frac{1}{10-1}[401 - \frac{(59)^2}{10}]}$$

$$= \sqrt{\frac{1}{9}[401 - 348.1]}$$

$$= \sqrt{\frac{52.9}{9}} = 2.42 \qquad \text{(point estimate)}$$

(iii) Point estimate of the standard error $\sigma_{\bar{x}}$ is ;

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \text{(estimator)}$$

$$= \frac{2.42}{\sqrt{10}} = 0.77 \qquad \text{(point estimate)}$$

(iv) Point estimate of the population proportion p is;

$$\hat{p} = \frac{X}{n} \qquad \text{(estimator)}$$

$$= \frac{5}{10} = 0.5 \qquad \text{(point estimate)}$$

6.2.2 Properties of a good point estimator

In point estimation the unknown value of a parameter is estimated by a single number which is quite risky job. This single value may or may not be equal to the true value of the parameter. The closeness of the estimate to the parameter value depends on random sample and the estimator. An ideal estimator is one which gives exactly the correct value of the parameter but such estimator does not exist in general. Hence to search a point estimate close to the parameter, it is necessary that the choice of one appropriate estimator in a given circumstance should be made on the basis of certain properties, called criteria for a good point estimator. A good point estimator is one which is unbiased, consistent,

efficient and sufficient. Only two properties, unbiasedness and efficiency are discussed here in detail.

6.2.3 Unbiasedness

It is not possible for an estimator to obtain correct estimate from each and every sample, even though if samples are drawn randomly. However, if this estimator on the average (i.e. considering all possible samples), give an estimate equal to the population parameter then this property of the estimator is called unbiasedness.

Definition of unbiased estimator:

An estimator is said to be unbiased if the mean of its sampling distribution is equal to the true value of the parameter, otherwise, the estimator is said to be biased. For example, if $\hat{\theta}$ is a point estimator of the parameter θ and $E(\hat{\theta})=\theta$, then $\hat{\theta}$ is called an unbiased estimator of θ. If $E(\hat{\theta})\neq\theta$ means that $\hat{\theta}$ is a biased estimator of θ. Further, if $E(\hat{\theta})>\theta$ means $\hat{\theta}$ is positively biased. If $E(\hat{\theta})<\theta$ means that $\hat{\theta}$ is negatively biased. Remember that in previous unit we have shown in practical examples that $E(\bar{X})=\mu_{\bar{x}}=\mu$. This implies that $\bar{X}=\frac{\sum x}{n}$ is an unbiased estimator for μ. Similarly, $E(\hat{p})=p$ means that $\hat{p}=\frac{X}{n}$ is an unbiased estimator of p. The bias of an estimator $\hat{\theta}$ is given by $B(\hat{\theta})=[E(\hat{\theta})-\theta]$.

Example 6.2

Show that sample mean $\bar{X}$ is an unbiased estimator of the population mean μ.

Proof:

Consider a random sample $X_1,\cdots,X_n$ from a normal population having mean μ and variance σ^2, then by definition,

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Taking expectation on both sides to have

$$E(\bar{X})=\frac{1}{n}E\left[\sum_{i=1}^{n}X_i\right]$$

$$=\frac{1}{n}[E(X_1)+E(X_2)+\cdots+E(X_n)]$$

$$=\frac{1}{n}(\mu+\mu+\cdots+\mu)$$ (As X_i's are drawn from a population having mean μ)

$$=\frac{n\mu}{n}=\mu$$ This implies that $\bar{X}$ is an unbiased estimator of μ.

Example 6.3

Show that sample proportion $\hat{p}$ is an unbiased estimator of population proportion p.

Proof:

By definition $\hat{p}=\frac{X}{n}$

$$E(\hat{p})=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)$$

$=\frac{1}{n}(np)$ (As X is Binomial random variable having mean np)

$=p$. It means that $\hat{p}$ is an unbiased estimator of p.

Example 6.4

Show that sample variance $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ is an unbiased estimator of population variance σ^2.

Proof:

By definition $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$(n-1)\, s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$

$= \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2$ i.e. μ is added and subtracted

$= \sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2$

$= \sum_{i=1}^{n}[(X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu)]$

$= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu)$

$= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)n(\bar{X} - \mu)$ As $\sum_{i=1}^{n}(X_i - \mu) = n(\bar{X} - \mu)$

$= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2n(\bar{X} - \mu)^2$

$= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$

Now taking expectation on both sides

$(n-1)E(s^2) = nE(X_i - \mu)^2 - nE(\bar{X} - \mu)^2$

$= n\sigma^2 - n\sigma_{\bar{x}}^2 \quad As\ \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

$= (n-1)\sigma^2$

$E(s^2) = \sigma^2$

This implies that $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of σ^2.

Example 6.5

Draw all possible samples of size 2 with replacement from a population having elements 4, 8, 12, 16 and show that;

i) $E(\bar{X}) = \mu$ *i.e.* $\bar{X}$ is an unbiased estimator of μ.

ii) $E(s^2) = \sigma^2$ i.e. $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of σ^2.

iii) $E(S^2) \neq \sigma^2$ *i.e* $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator of σ^2.

Solution:

Given 4, 8, 12, 16; $N = 4$; and $n = 2$

Population mean $\mu = \frac{\sum X}{N} = \frac{4+8+12+16}{4} = \frac{40}{4} = 10$

Population variance $\sigma^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{480}{4} - \left(\frac{40}{4}\right)^2 = 20$

Since sampling is with replacement therefore possible samples are $N^n = 4^2 = 16$

S.No	Samples	$\bar{X}$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2$
1	(4,4)	4	0	0
2	(4,8)	6	8	4
3	(4,12)	8	32	16
4	(4,16)	10	72	36
5	(8,4)	6	8	4
6	(8,8)	8	0	0
7	(8,12)	10	8	4
8	(8,16)	12	32	16
9	(12,4)	8	32	16
10	(12,8)	10	8	4
11	(12,12)	12	0	0
12	(12,16)	14	8	4
13	(16,4)	10	72	36
14	(16,8)	12	32	16
15	(16,12)	14	8	4
16	(16,16)	16	0	0

i) The sampling distribution of $\bar{X}$

$\bar{X}$	Tally bar	f	$f(\bar{x})$	$\bar{X}f(\bar{x})$
4	I	1	1/16	4/16
6	II	2	2/16	12/16
8	III	3	3/16	24/16
10	IIII	4	4/16	40/16
12	III	3	3/16	36/16
14	II	2	2/16	28/16
16	I	1	1/16	16/16
Total	-	16	1	160/16

$$E(\bar{X}) = \mu_{\bar{x}} = \sum \bar{x} f(\bar{x}) = \frac{160}{16} = 10 = \mu$$

As mean of the sampling distribution of $\bar{X}$ is equal to the population mean μ, so by definition of unbiasedness we say that $\bar{X}$ is unbiased estimator of μ.

(ii) The sampling distribution of s^2

s^2	Tally bar	f	$f(s^2)$	$s^2 f(s^2)$
0	IIII	4	4/16	0
8	IIII I	6	6/16	48/16
32	IIII	4	4/16	128/16
72	II	2	2/16	144/116
Total		16	1	320/16

$$E(s^2) = \mu_{s^2} = \sum s^2 f(s^2) = \frac{320}{16} = 20 = \sigma^2$$

Hence $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$ is an unbiased estimator of σ^2.

(iii) The sampling distribution of S^2

S^2	Tally bar	f	$f(S^2)$	$S^2 f(S^2)$
0	IIII	4	4/16	0
4	IIII I	6	6/16	24/16
16	IIII	4	4/16	64/16
36	II	2	2/16	72/16
Total		16	1	160/16

$$E(S^2)=\sum S^2 f(S^2)=\frac{160}{16}=10$$

$$E(S^2)\neq\sigma^2$$

Hence $S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ is a biased estimator of σ^2.

Example 6.6

Draw all possible samples of size 2 with replacement from a population consisting of units 1,2,5,6 and find the proportion of even numbers in the samples.

Construct the sampling distribution of sample proportion $\hat{p}$ and check that $E(\hat{p})=p$ i.e. $\hat{p}$ is an unbiased estimator of p.

Solution:

Given 1, 2, 5, 6 $\quad N=4 \quad n=2$

Population proportion $p=\frac{X}{N}=\frac{2}{4}=0.5$

Possible samples of size $n = 2$ in S.W.R. case are $N^n = 4^2 = 16$ which are shown in the table.

S.No.	Samples	$\hat{p}$
1	1,1	0
2	1,2	1/2
3	1,5	0
4	1,6	1/2
5	2,1	1/2
6	2,2	1
7	2,5	1/2
8	2,6	1
9	5,1	0
10	5,2	1/2
11	5,5	0
12	5,6	1/2
13	6,1	1/2
14	6,2	1
15	6,5	1/2
16	6,6	1

The sampling distribution of sample proportions $\hat{p}$

$\hat{p}$	Tally bar	f	$f(\hat{p})$	$\hat{p}f(\hat{p})$
0	IIII	4	4/16	0
1/2	IIII III	8	8/16	4/16
1	IIII	4	4/16	4/16
Total		16		8/16

$$E(\hat{p})=\mu_{\hat{p}}=\sum \hat{p}\, f(\hat{p})=\frac{8}{16}=0.5=p.$$

It means that $\hat{p}=\frac{X}{n}$ is an unbiased estimator of population proportion p.

6.2.4 Efficiency

If there are two estimators, both possessing the property of unbiasedness which can be used for the estimation of a parameter, it is difficult for us to choose the best one between them. The efficiency property decides to prefer the one which has minimum variance. Hence efficiency is a selection criterion for efficient estimator between unbiasead estimators.

Definition of efficient estimator:

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the same parameter θ and $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$, then $\hat{\theta}_1$ is efficient estimator than $\hat{\theta}_2$.

Efficiency is generally expressed in relative terms as:

$$R.E=\frac{Variance\ of\ efficient\ estimator}{Variance\ of\ other\ estimator}$$

If $0 \le R.E < 1 \Rightarrow \hat{\theta}_1$ is efficient

If $R.E > 1 \Rightarrow \hat{\theta}_2$ is efficient.

If R.E = 1 $\Rightarrow$ both estimators are equally efficient.

For example, in case of normal distribution both sample mean and sample median are unbiased estimators for μ. However $Var(\bar{X}) = \frac{\sigma^2}{n}$ and $Var(\text{median}) = \frac{\pi\sigma^2}{2n}$. Now comparing these estimators by relative efficiency criteria as:

$$R.E = \frac{Var(\bar{X})}{Var(median)} = \frac{\sigma^2/n}{\pi\sigma^2/2n} = \frac{\sigma^2}{n}.\frac{2n}{\pi\sigma^2} = \frac{2}{\pi}$$

$$= \frac{2}{22/7} = \frac{14}{22} = 0.64 < 1$$

This implies that $\bar{X}$ is more efficient than median for μ. Thus, an unbiased estimator which has minimum variance is called the best or the most efficient estimator.

Example 6.7

A random sample X_1, X_2, X_3 is drawn from a normal population having mean μ and variance σ^2. Let $\hat{\theta}_1 = \frac{X_1 + 2X_2 + X_3}{4}$ and $\hat{\theta}_2 = \frac{X_1 + X_2 + X_3}{3}$ are the estimators for μ.

i) Which of the estimators is unbiased?

ii) Show that $\hat{\theta}_1$ is efficient estimator for μ than $\hat{\theta}_2$.

Solution:

i) To check unbiasedness, we first consider:

$$\hat{\theta}_1 = \frac{X_1 + 2X_2 + X_3}{4}$$

Take expectation on both sides

$$E(\hat{\theta}_1) = \frac{1}{4}E[X_1 + 2X_2 + X_3]$$

$$= \frac{1}{4}[E(X_1) + 2E(X_2) + E(X_3)]$$

$$= \frac{1}{4}[\mu + 2\mu + \mu]$$, As all X_i's have same mean μ

$$= \frac{4\mu}{4} = \mu$$ Hence $\hat{\theta}_1$ is unbiased estimator of μ.

Now consider the second estimator:

$$\hat{\theta}_2 = \frac{X_1 + X_2 + X_3}{3}.$$

$$E(\hat{\theta}_2) = \frac{1}{3}[E(X_1) + E(X_2) + E(X_3)]$$

$$= \frac{1}{3}[\mu + \mu + \mu]$$

$$= \frac{3\mu}{3} = \mu$$ Thus, θ_2 is also an unbiased estimator of μ.

(ii) To check the efficiency, we have first compute variances for both estimators.

$$Var(\hat{\theta}_1) = Var\left[\frac{X_1 + 2X_2 + X_3}{4}\right]$$

$$= \frac{1}{16}[Var(X_1) + 4Var(X_2) + Var(X_3)]$$

$$= \frac{1}{16}[\sigma^2 + 4\sigma^2 + \sigma^2]$$, As X_i's are drawn from the population having variance σ^2.

$$= \frac{6\sigma^2}{16} = \frac{3\sigma^2}{8}.$$

Similarly $Var(\hat{\theta}_2) = Var\left[\frac{X_1 + X_2 + X_3}{3}\right]$

$$=\frac{1}{9}[Var(X_1)+Var(X_2)+Var(X_3)]$$

$$=\frac{1}{9}[\sigma^2+\sigma^2+\sigma^2] \quad =\sigma^2/3$$

Now compare the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ as;

$$R.E=\frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)}=\frac{3\sigma^2/8}{\sigma^2/3}=\frac{3\sigma^2}{8}\times\frac{3}{\sigma^2}=\frac{9}{8}=1.125>1\Rightarrow\hat{\theta}_2 \text{ is efficient than } \hat{\theta}_1$$

Thus relative efficiency criteria shows that the estimator $\hat{\theta}_2$ is efficient than $\hat{\theta}_1$. It means that $\text{var}(\hat{\theta}_2)<\text{var}(\hat{\theta}_1)$ and hence $\hat{\theta}_2$ is the best estimator for the parameter μ.

6.2.5 Pooled estimator from two samples

Parameters can also be estimated by estimators which are obtained by pooling (combining) the estimators computed from two or more random samples drawn from the same population, such an estimator is called pooled estimator. For example two random samples of sizes n_1 *and* n_2 have means $\bar{x}_1, \bar{x}_2$ and variances s_1^2, s_2^2 respectively, then combined mean given as $\bar{x}_c=\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$ is pooled estimator of μ and combined variance $s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ is pooled variance estimate of σ^2. Similarly, if random samples of sizes n_1, n_2 are drawn from a binomial population having sample proportions $\hat{p}_1$, $\hat{p}_2$, then $\hat{p}_c=\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$ is pooled estimator for p. If needed, pooled estimators from 3 or more samples can be obtained to extend in a similar way.

Example 6.8

Consider the following two samples

Sample A:	1	4	8	5
Sample B:	3	0	9	4

Compute pooled estimate for (i) population mean μ (ii) population variance σ^2 and (iii) proportion of odd numbers in the population.

Solution:

(i) $\bar{x}_1=\frac{1+4+8+5}{4}=\frac{18}{4}=4.5$ $\quad \bar{x}_2=\frac{3+0+9+4}{4}=\frac{16}{4}=4$

$\therefore \bar{x}_c=\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}=\frac{4(4.5)+4(4)}{4+4}=\frac{18+16}{8}=\frac{34}{8}=4.25$ is the pooled estimate of μ.

(ii) $s_1^2=\frac{1}{n_1-1}\left[\sum x_1^2-\frac{(\sum x_1)^2}{n_1}\right]=\frac{1}{4-1}\left[106-\frac{(18)^2}{4}\right]=\frac{1}{3}[25]=8.333$

$s_2^2=\frac{1}{n_{2-1}}\left[\sum x_2^2-\frac{(\sum x_2)^2}{n_2}\right]=\frac{1}{4-1}\left[106-\frac{(16)^2}{4}\right]=\frac{1}{3}[42]=14$

$\therefore s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}=\frac{(4-1)(8.333)+(4-1)(14)}{4+4-2}$

$=\frac{24.999+42}{6}=\frac{66.999}{6}=11.17$ is the pooled estimate of σ^2.

(iii) Proportion of odd numbers in sample A: $\hat{p}_1=\frac{2}{4}=0.5$

Proportion of odd numbers in sample B: $\hat{p}_2=\frac{2}{4}=0.5$

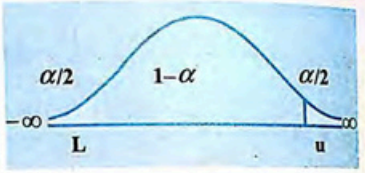
$\therefore \hat{p}_c=\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}=\frac{4(0.5)+4(0.5)}{4+4}=\frac{2+2}{8}=0.5$ is the pooled estimate of p.

6.3 Interval estimation

This process provides an interval of values calculated from sample data, as an estimate for the unknown population parameter. Interval estimate has a high probability of containing the parameter of interest.

6.3.1 Confidence interval

An interval which is constructed from sample observations in such a way that it has a high probability of containing the unknown value of parameter is called confidence interval. For example if L and U are two statistics, then the confidence interval for the parameter θ is given as:



$$P[L<\theta<U]=1-\alpha$$

▪ Main points about confidence interval

i. The end points that bound a confidence interval i.e. L and U are called critical values or confidence limits where L is the lower confidence limit and U is the upper confidence limit.

ii. The region between L and U is called confidence interval or confidence region or acceptance region for the unknown parameter θ.

iii. The region beyond the acceptance region i.e. from $-\infty$ to L and U to ∞ is known as critical region or rejection region.

iv. The probability that the interval contains the parameter is called confidence coefficient or confidence level and is denoted by $(1-\alpha)$. It is also written as $100(1-\alpha)\%$.

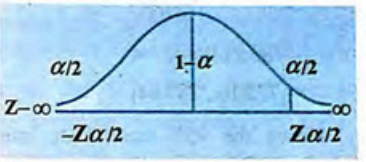
v. The probability that parameter lies in the rejection region is called significance level and is denoted by α.

vi. $(U-L)$ is called width or length of confidence interval which is a measure of precision for confidence interval.

vii. Precision means accuracy of the confidence interval. It can be increased either by increasing the sample size or decreasing the confidence coefficient.

6.3.2 Large sample confidence interval for population mean μ (when σ is known):

To find a $100(1-\alpha)\%$ large sample $(n \geq 30)$ confidence interval for population mean μ when σ is known, we begin with the statistic $Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ which has a standard normal distribution and make a probability statement as:



$$P[-Z_{\alpha/2} < Z < Z_{\alpha/2}] = 1-\alpha$$

Where $Z_{\alpha/2}$ means that probability or area to the right of $Z_{\alpha/2}$ is equal to $\alpha/2$. Similarly probability to the left of $-Z_{\alpha/2}$ is equal to $\alpha/2$ such that $\alpha/2+\alpha/2=\alpha$.

A $100(1-\alpha)\%$ confidence interval for μ when σ is known is given by;

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Note for particular sample use $\bar{x}$ instead of $\bar{X}$.

Example 6.9

An electrical firm manufactures light bulbs that have a length of life with mean μ and a standard deviation of 40 hours. If a random sample of 100 bulbs has an average life of 780 hours, find a 95% confidence interval for the population mean of all bulbs produced by this firm.

Solution:

A $100(1-\alpha)\%$ confidence interval for μ is $\bar{X} \pm Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

We are given $\sigma = 40$, $n = 100$, $\bar{x} = 780$, and confidence level $1-\alpha = 95\% = 0.95$ so $\alpha = 1-0.95 = 0.05$, $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

Now make inverse use of area table of standard normal distribution and search $0.5-0.025 = 0.475$ in the body of the table which correspond to $Z_{\alpha/2} = Z_{0.025} = 1.96$.

Putting values in the above interval estimator, we get

$$780 \pm 1.96 \frac{40}{\sqrt{100}}$$

$$780 \pm 1.96(4)$$

$$[772.16, 787.84]$$

Hence the 95% confidence interval for μ is from 772.16 to 787.84 hours.

Example 6.10

A random sample of size n = 400 selected without replacement from a population of size N = 2000 with $\sigma = 4$ gives $\bar{x} = 80$. Use this sample information to construct a 90% confidence interval for the mean of the population.

Solution:

A $100(1-\alpha)\%$ confidence interval for μ is $\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

Given that $N = 2000$, $n = 400$, $\sigma = 4$, $\bar{x} = 80$,

$1-\alpha = 0.90$, $\alpha = 0.10$, $\frac{\alpha}{2} = 0.05$.

From the area table we have $Z_{\alpha/2} = Z_{0.05} = 1.645$ corresponding to the probability $0.50 - 0.05 = 0.45$

Hence the 90% confidence interval for μ is

$$80 \pm 1.645 \frac{4}{\sqrt{400}} \sqrt{\frac{2000-400}{2000-1}}$$

or 80 ± 0.294

or $80 - 0.294$, $80 + 0.294$

or $[79.706, 80.294]$

6.3.3 Large sample confidence interval for population mean μ when σ is unknown

Practically, σ is usually not known but if $n \geq 30$ then, central limit theorem allows us to consider the sampling distribution of $\bar{X}$ as approximately normal with mean μ and S.E $= \frac{s}{\sqrt{n}}$ that is, σ is replaced by sample standard deviation s. An approximate $100(1-\alpha)\%$ confidence interval for μ in this case is given as;

$$\bar{X} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}.$$

Example 6.11

A scientist, interested in monitoring chemical contaminants in food, selected a random sample of $n = 50$ male adults. It was found that the average daily intake of dairy products was $\bar{x} = 756$ grams per day with a standard deviation of $s = 35$ grams per day. Construct a 95% confidence interval for the mean daily intake of dairy products for males.

Solution:

Here σ is unknown but n = 50 is large, therefore, $100(1-\alpha)\%$ confidence interval for μ is:

$$\bar{X} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$$

We have $1-\alpha = 0.95$, $\alpha = 0.05$, $Z_{\alpha/2} = Z_{0.025} = 1.96$

Hence the approximate 95% confidence interval for μ is

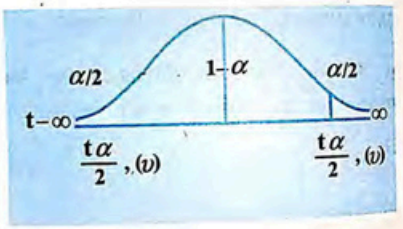
$756 \pm 1.96 \frac{35}{\sqrt{50}}$ or 756 ± 9.70

or $756-9.70$, $756+9.70$ or [746.30 , 765.70]

This means that the mean daily intake of dairy products for males varies from 746.30 to 765.70 grams per day.

6.3.4 Confidence interval for population mean μ when σ is unknown (small sample)

When σ is known and n is small, then the interval estimator $\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ will still be used for μ but if σ is unknown and $n < 30$, then the statistic $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ is used which has t-distribution with $v = n-1$ degrees of freedom and $s = \sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2}$.



Probability statement can be made as:

$$P\left[-t_{\alpha/2,(v)} < t < t_{\alpha/2,(v)}\right] = 1-\alpha$$

Putting value of t to have a $100(1-\alpha)\%$ confidence interval for μ when σ is unknown and n is small as;

$$\bar{X} \pm t_{\alpha/2,(v)} \frac{s}{\sqrt{n}}$$

- **Degrees of freedom**

The number of values in the sample that are independent and free to vary is called degrees of freedom e.g. if $x_1 + x_2 = 10$. we can give value only to x_1 or x_2 the second one will be automatically computed. Hence $d.f = 2-1 = 1$. Generally if we have $x_1 + x_2 + \ldots\ldots x_n = 10$ we can gave values to first n-1 scores freely and one score is restricted, so we have degrees freedom $d.f = n-1$

TABLE 6.1 [Critical values of student t-distribution]

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.204	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.146	2.650	3.012	13
14	1.345	14.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
∞	1.282	1.645	1.960	2.326	2.576	∞

Example 6.12

Find a 99% confidence interval for population mean when a random sample selected from the population gives the values 1.03, 1.01, 0.097, 1.04, 0.99, 0.98, 1.03, 1.01, 0.99.

Solution:

A $100(1-\alpha)\%$ confidence interval for μ is $\bar{X} \pm t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}}$

We have values 1.03, 1.01, 0.97, 1.04, 0.99, 0.98, 1.03, 1.01, 0.99.

Here $n=9$, $\Sigma x=9.05$, $\Sigma x^2=9.1051$, therefore,

$$\bar{x}=\frac{\Sigma x}{n}=\frac{9.05}{9}=1.0056,$$

$$s=\sqrt{\frac{1}{n-1}\left[\Sigma x^2-\left(\frac{\Sigma x}{n}\right)^2\right]}=\sqrt{\frac{1}{9-1}\left[9.1051-\frac{(9.05)^2}{9}\right]}=0.02245$$

$1-\alpha=0.99$ or $\alpha=0.01$

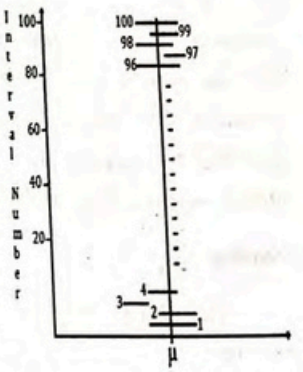
Now form t-table 6.1, we have $t_{\alpha/2,(n-1)}=t_{0.01/2,(9-1)}=t_{0.005,(8)}=3.355$

Putting values in expression for confidence interval, we get 99% confidence interval for μ as

$$1.0056 \pm 3.355\left(\frac{0.0245}{\sqrt{9}}\right)$$ or [1.0056 ± 0.0274], [0.9782 , 1.033]

6.3.5 Interpreting the confidence interval

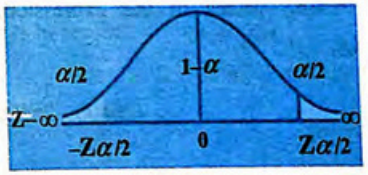
What does it mean to say we are 98% confidant that the true value of the population mean μ is within a given interval? It means that if we construct 100 such intervals for different samples, almost 98 out of 100 will contain the parameter and only 2 out of 100 will not contain the true value of the parameter. It can diagrammatically be shown as given in the Figure. Two of the intervals at serial number 3 and 97 do not contain the parameter μ while remaining 98 intervals contain the parameter. Remember that we cannot be absolutely sure that any one particular interval contains the mean μ.



6.3.6 Confidence interval estimate for the difference between two populations means when σ_1 and σ_2 are known

In this case $\bar{X}_1-\bar{X}_2$ is the point estimator of $(\mu_1-\mu_2)$ which in standard form can be written as;

$$Z=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}.$$



Probability statement for the random variable Z can be made as $P[-Z_{\alpha/2}<Z<Z_{\alpha/2}]=1-\alpha$.

The $100(1-\alpha)\%$ confidence interval for $(\mu_1-\mu_2)$ in this case is given by

$$(\bar{X}_1-\bar{X}_2)\pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Example 6.13

The wearing qualities of two types of automobile tires were compared by road – testing samples of $n_1 = n_2 = 100$ tires of each type. The test results are $\bar{x}_1 = 26400$ miles and $\bar{x}_2 = 25100$ miles. The standard deviations for the two types of populations are $\sigma_1 = 37.9473$ and $\sigma_2 = 44.2719$ respectively. Estimate $(\mu_1 - \mu_2)$, the difference in mean miles to wear out, using a 99% confidence interval.

Solution:

The $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ in this case is given by

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Given $n_1 = 100$ $\bar{x}_1 = 26400$ $\sigma_1 = 37.9473$ $\sigma_1^2 = 1440$

$n_2 = 100$ $\bar{x}_2 = 25100$ $\sigma_2 = 44.2719$ $\sigma_2^2 = 1960$

$1-\alpha = 0.99$ $\alpha = 0.01$ $\frac{\alpha}{2} = 0.005$

Hence $z_{\alpha/2} = 2.57$

Putting values in the expression, we have 99% confidence interval for $(\mu_1 - \mu_2)$ as;

$$(26400 - 25100) \pm 2.57\sqrt{\frac{1440}{100} + \frac{1960}{100}}$$

$$1300 \pm 2.57(5.831)$$

or $[1285.014 < (\mu_1 - \mu_2) < 1314.986]$

6.3.7 Confidence interval for $(\mu_1 - \mu_2)$ when σ_1 and σ_2 are unknown (large sample case)

When population variances are unknown but sample sizes n_1 and n_2 are large (≥ 30) then σ_1^2, σ_2^2 are estimated by sample variances S_1^2, S_2^2 respectively. The statistic Z in this case will be of the form

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The $100\,(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ is then

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Example 6.14

Intelligence test of two groups of boys and girls gave the following results:

Girls: $n_1 = 60$ mean $= 75$ $S.D = 8$

Boys: $n_2 = 100$ mean $= 73$ $S.D = 10$

Compute a 95% confidence interval for $(\mu_1 - \mu_2)$ where μ_1 is the mean score of girls and μ_2 is the mean score of boys in respective populations.

Solution:

Here σ_1, σ_2 are not given but n_1, n_2 are large, so $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ in this case is

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

We have $n_1 = 60$, $\bar{x}_1 = 75$, $S_1 = 8$

$n_2 = 100$, $\bar{x}_2 = 73$, $S_2 = 10$

$1-\alpha = 0.95$, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$

From area table of normal distribution, we have $Z_{\alpha/2} = Z_{0.025} = 1.96$

Hence the 95% confidence interval for $(\mu_1 - \mu_2)$ is:

$$(75-73) \pm 1.96 \sqrt{\frac{64}{60} + \frac{100}{100}}$$

$$2 \pm 1.96\ (1.44)$$

$$2 \pm 2.8224$$

$$(-0.8224\ ,\ 4.8224)$$

6.3.8 Confidence interval estimate for the difference between two normal populations means $(\mu_1 - \mu_2)$ when σ_1^2 , σ_2^2 are unknown (small sample case)

When population variances are not given and n_1 , n_2 are small (< 30) then we have to make another assumption i.e. $\sigma_1 = \sigma_2 = \sigma$ and this σ is estimated by unbiased pooled estimator. In this case the statistic $\bar{X}_1 - \bar{X}_2$ has the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom given below;

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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

A $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ in this case is given by

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,(n_1+n_2-2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 6.15

The following two samples were randomly selected from two normal populations for which $\sigma_1^2 = \sigma_2^2$ but unknown.

Sample-I	103	94	110	87	98		
Sample-II	97	82	123	92	175	88	118

Compute 90% confidence interval for the difference between the two population means.

Solution:

$100(1-\alpha)\%$ confidence Interval for $(\mu_1 - \mu_2)$ in this case is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,(n_1+n_2-2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Given that

$n_1 = 5$, $\Sigma x_1 = 492$, $\Sigma x_1^2 = 48718$

$n_2 = 7$, $\Sigma x_2 = 775$, $\Sigma x_2^2 = 92019$

Now $\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{492}{5} = 98.4$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{775}{7} = 110.71$$

$$(n_1-1)s_1^2 = \Sigma(x_1 - \bar{x}_1)^2 = \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1} = 48718 - \frac{(492)^2}{5} = 305.2$$

$$(n_2-1)s_2^2 = \Sigma(x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2} = 92019 - \frac{(775)^2}{7} = 6215.4286$$

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

$$= \sqrt{\frac{305.2+6215.4286}{5+7-2}} \quad = \sqrt{\frac{6520.6286}{10}} = 25.54$$

Here $1-\alpha = 0.90$, $\alpha = 0.10$

From $t-$table 6.1 we have $t_{\alpha/2,(n_1+n_2-2)} = t_{0.10/2,(5+7-2)} = t_{0.05,(10)} = 1.812$

Hence 90% confidence Interval for $(\mu_1 - \mu_2)$ is

$$(98.4-110.71) \pm 1.812(25.54)\sqrt{\frac{1}{5}+\frac{1}{7}}$$

$$-12.31 \pm 27.10$$

$$[-39.41 < \mu_1 - \mu_2 < 14.79]$$

6.3.9 Confidence interval estimate for population proportion p (large sample case)

When the sample size is large the sample proportion $\hat{p}$ has the sampling distribution which is approximately normal with mean p and $S.E = \sqrt{\frac{pq}{n}}$ i.e.

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \text{ as } n \to \infty$$

A $100(1-\alpha)\%$ confidence interval for p is written as;

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

If p is unknown then the confidence interval for p is computed by

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Example 6.16

A random sample of 200 persons from a city was interviewed and 50 of them were found to be literate. Calculate a 90% confidence interval for the proportion of literate persons in the city.

Solution:

A $100(1-\alpha)\%$ confidence interval for p (literate persons) is

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Given

$n = 200$, $X = 50$ (Number of literate persons), therefore

$$\hat{p} = \frac{X}{n} = \frac{50}{200} = 0.25, \quad \hat{q} = 1-\hat{p} = 1-0.25 = 0.75,$$

$1-\alpha = 0.90$, $\alpha = 0.10$. From area table of standard normal distribution, we have $Z_{\alpha/2} = Z_{\frac{0.10}{2}} = Z_{0.05} = 1.645$

Hence the 90% confidence interval for p is $0.25 \pm 1.645\sqrt{\frac{(0.25)(0.75)}{200}}$

or 0.25 ± 0.05 or $(0.2\ ,\ 0.3)$ or $(0.2 < p < 0.3)$

6.3.10 Large sample confidence interval estimate for the difference between two population proportions ($p_1 - p_2$)

A simple extension of the estimation of a binomial proportion p is the estimation of the difference between two binomial proportions. For sufficiently large sample sizes the sampling distribution of $\hat{p}_1 - \hat{p}_2$ can be approximated by a normal distribution i.e.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \sim N(0,1)$$

The approximate $100(1-\alpha)\%$ confidence interval for $(p_1 - p_2)$ is given by:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Example 6.17

Two pairs relieving drugs were compared each on independent samples of $n_1 = 1000$, $n_2 = 1000$ individuals. Out of these individuals 750 receiving drug-I and 800 receiving drug-II reported some pain relief. Construct a 90% confidence interval for the difference between population proportions.

Solution:

As p_1, p_2 (population proportions) are unknown, therefore $100(1-\alpha)\%$ confidence interval for $(p_1 - p_2)$ is $(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Given information are;

$$n_1 = 1000 \ , \quad X_1 = 750 \ , \ \hat{p}_1 = \frac{X_1}{n_1} = \frac{750}{1000} = 0.75$$

$$\hat{q}_1 = 1 - \hat{p}_1 = 1 - 0.75 = 0.25,$$

$$n_2 = 1000 \ , \quad X_2 = 800 \ , \ \hat{p}_2 = \frac{X_2}{n_2} = \frac{800}{100} = 0.80$$

$$\hat{q}_2 = 1 - \hat{p}_2 = 1 - 0.80 = 0.20,$$

$$1 - \alpha = 0.90 \ , \quad \alpha = 0.10 \ , \quad \frac{\alpha}{2} = 0.05.$$

From the area table of standard normal distribution, we have $Z_{\alpha/2} = Z_{0.05} = 1.645$.

Hence, the 90% confidence interval for $(p_1 - p_2)$ is

$$(0.75 - 0.80) \pm 1.645\sqrt{\frac{(0.75)(0.25)}{1000} + \frac{(0.80)(0.20)}{1000}}$$

$$-0.05 \pm 0.03$$

$$(-0.08 \ , \ -0.02)$$

Key points

- Statistical inference is a process of drawing conclusions (inferences) about the population on the basis of sample information from that population.
- Statistical estimation is a process by which the unknown value of a parameter is obtained from the sample observations.
- When a specific value is obtained from sample observations and is used to estimate the unknown value of the parameter, the process is called point estimation.
- When a range of values is obtained from sample observations within which the unknown value of parameter is believed to lie, the process is called interval estimation.
- A rule, usually expressed as a formula that tells us how to calculate an estimate from the sample data is called an estimator and the resulting number is called an estimate.
- An estimator is said to be unbiased if the mean of its sampling distribution is equal to the true value of the parameter, otherwise the estimator is said to be biased
- If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the same parameter θ and $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$ then $\hat{\theta}_1$ is called efficient estimator than $\hat{\theta}_2$.
- An estimator which is linear, unbiased and has minimum variances among a group of linear unbiased estimators is called best linear unbiased estimator or BLUE.
- A rule for calculating two numbers to create an interval which has a high probability of containing the parameter of interest is the confidence interval.
- The region between L and U is called confidence interval or confidence region or acceptance region for the unknown parameter θ.
- The region beyond the acceptance region i.e. from $-\infty$ to L and U to ∞ is known as critical region or rejection region.
- The probability that parameter lies in the rejection region is called significant level and is denoted by α.

Exercise

6.1 Read the following statements carefully and indicate which statement is true or false.

i. Statistical inference means making confidence interval for the parameter.

ii. Parameters are constant and statistics are random variables.

iii. Inference regarding population parameters can be done in three ways.

iv. A statistic will always be an unbiased estimator if the sample itself is chosen without bias.

v. $s^2 = \frac{1}{n-1}\Sigma(X-\bar{X})^2$ is an unbiased estimator of σ^2.

vi. In estimation procedure we estimate the value of a statistic.

vii. A 99% confidence interval will be wider than a 95% confidence interval constructed from the same data.

viii. The precision of the confidence interval will increase by increasing the sample size.

ix. Confidence limits will vary from sample to sample.

x. t-distribution is used for interval estimation of μ when σ is known and n is large.

6.2 Fill in the blanks.

i. An estimator is itself a _____.

ii. A value of an estimator is called an _____.

iii. If $x_1, x_2 ... x_n$ be a random sample, the expression $\bar{x} = \frac{\Sigma x}{n}$ is an _____.

iv. A single value of an estimator for a population parameter is called its _____estimate.

v. If expected value of an estimator $\hat{\theta}$ is equal to the value of the parameter θ, then $\hat{\theta}$ is said to be an _____ estimator of θ.

vi. If two estimators T_1 and T_2 such that $Var(T_1) < Var(T_2)$, then T_1 is called _____estimator than T_2.

vii. An interval estimate with _____ interval is best.

viii. The precision of a confidence interval increases by _____ the sample size.

ix. Estimation has _____ types.

x. $(1-\alpha)$ is known as _____.

6.3 Choose the Correct answer:

i. Estimate and estimator are:

(a) synonyms (b) different

(c) related to population (d) formulae

ii. Bias of an estimator can be:

(a) positive (b) negative

(c) either positive or negative (d) always zero

iii. If $\hat{p} = 0.5$, then 0.5 is called ;

(a) estimator (b) estimate

(c) interval (d) all of above

iv. Estimation has _____ types.

(a) 3 (b) 4

(c) 2 (d) 5

v. α is called ;

(a) level of significance (b) confidence level

(c) confidence coefficient (d) all of above

vi. The probability statement $P[-Z_{\alpha/2} < Z < Z_{\alpha/2}] = ?$

(a) α (b) β

(c) $1-\beta$ (d) $1-\alpha$

vii. If the average value of an estimator is equal to the true value of the parameter, the property is called;

(a) efficiency (b) sufficiency

(c) consistency (d) unbiasedness

viii. Statistical inference makes inferences about

(a) Sample (b) population

(c) both population and sample (d) estimator

ix. $\left[\hat{p}-Z_{\alpha/2}\sqrt{\frac{pq}{n}} < p < \hat{p}+Z_{\alpha/2}\sqrt{\frac{pq}{n}}\right]$ is an interval estimate for

(a) mean (b) variance

(c) proportion (d) S.D

x. A 90% confidence interval for the population mean is of the from

(a) $\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}}$ (b) $\bar{x} \pm 1.28\frac{\sigma}{\sqrt{n}}$

(c) $\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$ (d) $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

6.4 What do you understand by estimation? Differentiate between an estimator and an estimate.

6.5 Describe the following:

i) Statistical inference ii) Types of estimates.

6.6 Distinguish between

i) Point estimate and interval estimate

ii) Estimator and estimate.

6.7 Define and discuss with examples the following properties of a point estimator;

i) unbiasedness. ii) efficiency.

6.8 What do you mean by bias? What are the factors which introduce bias?

6.9 Given a random sample 4, 3, 7, 8, 4, 10, 5, 5, 4, 9 Compute point estimate for (i) population mean (ii) population variance (iii) S.E of the mean.

6.10 If random samples of size n = 2 are drawn with replacement from population having values 10, 11, 13, 15, 16, 19.

Show that;

i) Sample mean is an unbiased estimator of population mean

ii) Sample variance $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$ is unbiased estimator of the population variance.

6.11 A random sample of size 3 i.e. x_1, x_2, x_3 is drawn from a normal population having mean μ and variance σ^2. The following two statistics $T_1 = \frac{x_1+2x_2+x_3}{4}$, $T_2 = \frac{x_1+x_2+x_3}{3}$ are taken for μ.

i) Which of the above is unbiased?

ii) Which of the above is most efficient?

6.12 Describe the concept of confidence interval estimation.

6.13 Define the following terms:

i) confidence limits, ii) confidence coefficient iii) level of significance, iv) precision of a confidence interval.

6.14 A random sample 12, 9, 14, 10, 12, 07, 13, 11 is drawn from a normal population whose $\sigma = 2$. Compute a 90% confidence interval for the mean of this normal population.

6.15 A sample of 100 chocolate bars is taken at random from a large shipment have an average $\bar{x} = 0.8$ pound with a standard deviation of $S = 0.1$ pound. Find a 99% confidence interval for the mean weight (μ) of chocolate bars for the entire shipment.

6.16 An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 42 hours. If a random sample of 49 bulbs has an average life of 800 hours. Find a 95% confidence interval for the population mean of all bulbs produced by this firm.

6.17 Given the sample 2.3, −0.2, −0.4, −0.9. Compute a 90% confidence interval for the mean of a normal population with σ=3.

6.18 The heights of a random sample of 100 college students showed a mean height of 64 inches. If standard deviation of the height distribution of the population is 3 inches, find a 95% confidence interval for the mean height of the population.

6.19 If $n=50$, $\Sigma x=2163$, $\Sigma x^2=144949.6$. Compute a 99% confidence interval for μ.

6.20 A random sample of size $n_1=36$ taken from a normal population with a variance $\sigma_1^2=9$ has mean $\overline{x}_1=75$. A second random sample of size $n_2=25$ taken from a different population with a variance $\sigma_2^2=25$ has a mean $\overline{x}_2=70$. Find a 98% confidence interval for $\mu_1-\mu_2$.

6.21 The number of accidents per day in two cities was observed and the following information was obtained:

Description	city-A	city-B
Numbers of days	144	100
Mean number of accidents	4.5	5.4
Standard deviation	1.2	1.5

Estimate 95% confidence interval for the difference between the mean accidents of the two cities.

6.22 When it is appropriate to use t-distribution instead of the Z-distribution to construct confidence interval for the population mean or difference of population means?

6.23 A sample of 10 measurements of the diameter of a spare gave a mean $\overline{x}=4.38$ inches and a standard deviation s = 0.06 inches. Find a 95% confidence interval for the actual diameter.

6.24 A random sample of 12 ball bearings has weights in grams as 31.4, 33.1, 35.9, 34.7, 33.4, 34.5, 35, 32.5, 36.9, 36.4, 35.8, 33.2. Find a 90% confidence interval for the mean weight of the population from which these weights were drawn.

6.25 The following summary statistics were recorded from independent random samples drawn from two populations:

	n	$\overline{x}$	s^2
Sample A	10	74	60
Sample B	13	81	40

Construct a 99% confidence interval for $\mu_A-\mu_B$

6.26 Find a 95% confidence interval for p if 24 heads are obtained in 40 tosses of a coin.

6.27 In a survey carried out in a large city 170 housewives out of a random sample of 250 preferred Lipton brand of tea. Find a 95% confidence interval for the percentage of all housewives in the city preferring Lipton brand of tea.

6.28 Independent random samples of $n_1=800$ and $n_2=640$ observations were selected from binomial populations 1 and 2, and $X_1=337$, $X_2=374$ successes were observed. Find a 90% confidence interval for the difference (p_1-p_2) in the two population proportions. Interpret the interval.

Unit - 7

Hypothesis Testing

After studying this unit, the students will be able to

- Describe statistical hypothesis and hypothesis testing.
- Differentiate between null and alternative hypothesis, simple and composite hypothesis.
- Formulate null and alternative hypothesis.
- Recognize the elements involved in hypothesis testing: test statistic, rejection and non-rejection regions, critical values, one tailed test, two tailed test, type-I and type-II errors, level of significance, decision rule and conclusion.
- Apply the test of hypothesis about mean of a normal population (known and unknown standard deviation)
- Apply the test of hypothesis about population proportion (large sample).
- Apply the test of hypothesis about the difference between means of two normal populations (known/unknown standard deviations).
- Apply the test of hypothesis about the difference between proportions of two populations (large samples).

7.1 Introduction

The second important phase of statistical inference is the hypothesis testing. Here decisions are made about the population parameters on the basis of sample information. For example, hypothesis testing procedure can decide whether (i) a new medicine is really effective in curing a particular disease (ii) one educational procedure is better than other etc. The difference between estimation and hypothesis testing is that in estimation parameter values are unknown and are obtained from sample information while in hypothesis testing parameters values are hypothesized and are checked on the basis of sample information whether they are true or false.

Hypothesis is an unproved claim or assertion or assumption which acts as a starting point in a research irrespective of its probable truthfulness or falsity. Hypothesis may be non-statistical or statistical. A statistical hypothesis is a testable claim about one or more parameters of empirical distributions.

7.1.1 Hypothesis testing

A statistical method that uses sample data to accept or reject a hypothesis about a parameter is called hypothesis testing.

7.1.2 Principle steps involved in hypothesis testing

Hypothesis testing procedure mainly involves the following six steps:

Step i. State the null and alternative hypothesis.

Step ii. Choose an appropriate level of significance.

Step iii. Decide about test statistic.

Step iv. Compute test statistic value.

Step v. Obtain critical values for test statistic from respective tables.

Step vi. Make a decision.

7.1.3 Definitions of key terms

- **Null hypothesis**

It is a claim about parameter. It is tested for possible rejection under the assumption that it is true. It is denoted by H_0.

- **Alternative hypothesis**

It is also a claim about parameter which is accepted if H_0 is rejected. It is denoted by H_1. It may be directional or non-directional.

- **Simple hypothesis**

When all parameters of a distribution are well specified under the hypothesis it is called simple hypothesis. For example in case of normal distribution if H_0: $\mu = 5, \sigma^2 = 2.5$, it is simple hypothesis because all parameters values are well specified under H_0.

- **Composite hypothesis**

When all parameters of a distribution are not well specified under the hypothesis it is called composite hypothesis. For example for normal distribution H_0: $\mu = 5$, $\sigma^2 \leq 2.5$ or H_0: $\mu < 5$, $\sigma^2 = 2.5$ or H_0: $\mu = 5$, σ unknown etc. are composite hypothesis because in the first two cases values are not well specified and in the third case all parameters are not considered. The concept of simple and composite hypothesis is applicable to both H_0 and H_1.

- **Type-I and Type-II errors**

Two types of errors are possible in making a decision about a hypothesis.

(1) Type-I error

"To reject a null hypothesis when it is true, it is called type-I error"

Examples

i. If a medicine is given to a few patients of a particular disease to cure them and the medicine is curing the disease, but it is claimed that it has no effect or has an adverse effect and hence discontinued. This is Type-I error.
ii. To fail an intelligent student.
iii. To punish an innocent person.
iv. When a good player is not played in the cricket match.
v. To reject a good item.

The probability of committing type-I error is denoted by α i.e. $\alpha = p$(reject H_0/H_0 is true)

(2) Type-II error

"To accept a null hypothesis when it is false, it is called type-II error."

Examples

i. If the medicine has adverse effect and is claimed to have good effect and the treatment is continued. This is type-II error.
ii. To pass a dull student.
iii. To release a guilty person.
iv. To allow a non-deserving player in the cricket match.
v. To accept a bad item.

The probability of committing type-II error is denoted by β i.e. $\beta = p$(accept H_0/H_0 is false)

Remember that

(i) Type-I error is more serious than type-II error.

(ii) α and β have an inverse relation i.e. if one increases the other decreases and vice-versa.

(iii) Both α and β can be decreased by increasing n.

• Significance level

The probability of type-I error which we are ready to tolerate in making decision about H_0 is called significance level (S.L). It is denoted by α i.e. $\alpha = P(\text{reject } H_0 / H_0 \text{ is true})$. The higher the significance level we use for testing a hypothesis, the higher the probability of rejecting H_0 when it is true. Usually, it is suggested to keep $\alpha \leq 0.05$.

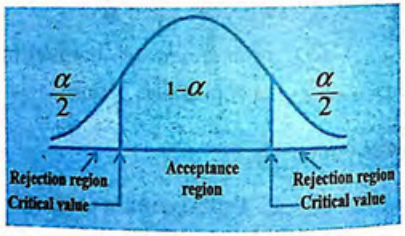
• Test statistic and test of significance

A function of sample data is called statistic. When a statistic is used to test a hypothesis, it is called test statistic or decision rule. It helps us in the decision whether to accept or reject H_0. The choice of test statistic depends on the hypothesis under question. Commonly used test statistics are Z, t, χ^2 and F.

The value of test statistic is said to be statistically significant if it falls in the rejection region and as a result H_0 is rejected. On the other hand, value of test statistic is said to be statistically insignificant if it falls in the acceptance region and as a result H_0 is accepted.

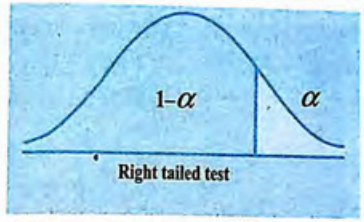
• Acceptance and rejection regions

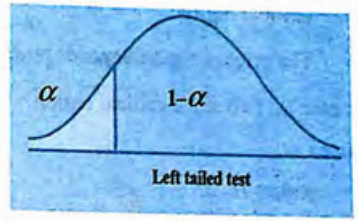
All possible values of a test statistic are divided into two mutually exclusive groups. The group of values which lead to the acceptance of H_0 is called acceptance region (AR) for the test. The group of values which would lead us to rejection of H_0 is called rejection region (RR) or critical region for the test. The values which separate the (AR) and (RR) are called critical values and are obtained from the tables of respective test statistics.



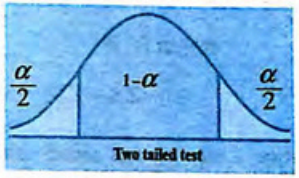
• One-tailed and two-tailed test

Three types of symbols [>, <, ≠] are used in the alternative hypothesis H_1. When the symbols [>, <] are used in H_1, it means that the whole rejection region of size α lies only on one tail of the concerned sampling distribution as shown below,





It is called one-sided rejection region and the test is called one-sided or one tailed test. On the other hand when the symbol [≠] is used in H_1, it means that the whole rejection region of size α is equally divided; $\alpha/2$ lies in the right tail and $\alpha/2$ lies in the left tail of the concerned sampling distribution, as shown in the figure. It is called two sided rejection region and the test is called two-tailed test.



7.2 General procedure for testing hypothesis about Mean (μ) of a Normal Population [when σ^2 is known]

i. Generally H_0 and H_1 can be set in three different possible forms as follow:

a)	b)	c)
$H_0: \mu = \mu_0$	$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$
[Two tailed test]	[Left tailed test]	[Right tailed test]

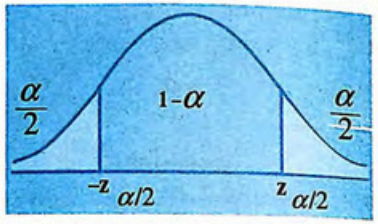
ii. $\alpha = 0.01$ or 0.05 etc.

iii. Test statistic to be used is $Z = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, which follows N(0,1) when H_0 is true.

iv. Calculation of Z value from the given data.

v. Critical region:

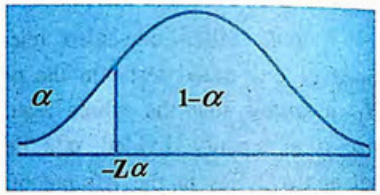
The critical region depends on H_1.



For case (a) two sided critical region is used as;

Reject H_0 if

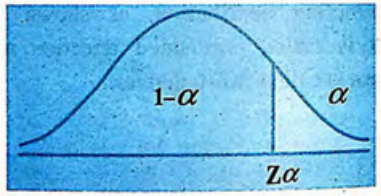
$$Z \le -Z_{\alpha/2} \text{ or } Z \ge Z_{\alpha/2}.$$



For case (b) one sided critical region is used as;

Reject H_0 if

$$Z \le -Z_{\alpha}$$



For Case (c) one sided critical region is used as;

Reject H_0 if

$$Z \ge Z_{\alpha}$$

Where Z is computed value of Z-test and $Z_{\alpha/2}$ or Z_{α} is table value.

vi. Conclusion:

If the computed value of Z falls in the acceptance region, accept H_0, otherwise reject H_0. Remember that acceptance of a hypothesis does not mean that it is really true.

We may interpret it as that sample data is in the support of H_0 that is why we are accepting H_0.

Example 7.1

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a mean of 812 hours and a standard deviation of 40 hours. Test the hypothesis that $\mu = 812$ hours against the alternative $\mu \neq 812$ hours if a random sample of 36 bulbs has an average life of 800 hours. Use a 5% level of significance.

Solution:

i. Null hypothesis H_0: $\mu = 812$

Alternative hypothesis H_1: $\mu \neq 812$

ii. Significance level $\alpha = 0.05$

iii. Test statistic to be used here is;

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

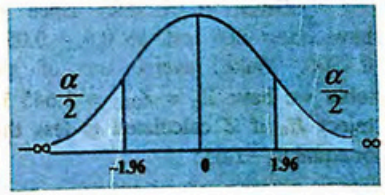
iv. Calculation: Given n = 36, $\bar{X} = 800$, $\sigma = 40$

Hence $Z = \dfrac{800-812}{40/\sqrt{36}} = \dfrac{-12}{40/6} = \dfrac{-12\times 6}{40} = -1.8$

v. Critical region:

Here the test is two sided, so $\alpha/2 = 0.025$ Now $0.5 - 0.025 = 0.4750$ search this value in the body of the area table of standard normal distribution which correspond to $Z_{\alpha/2} = Z_{0.05} = 1.96$.

Thus reject H_0 if $Z \le -1.96$ or $Z \ge 1.96$.



vi. Conclusion:

As computed value of $Z = -1.8$ falls in the acceptance region, therefore, we accept H_0 i.e. mean life of the bulbs produced by this firm is equal to 812 hours.

Example 7.2

A random sample of 25 hens from a normal population showed that the average laying is 250 eggs per year. The company claims that the average laying is 285 eggs per year with a standard deviation of 25 eggs per year. Test the claim of the company against the alternative that average laying is less than 285 eggs at $\alpha = 0.05$

Solution:

i. $H_0: \mu = 285$

$H_1: \mu < 285$

ii. $\alpha = 0.05$

iii. Test statistic: As $n = 25$ is small but σ is known, therefore, we use test statistic for μ as $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

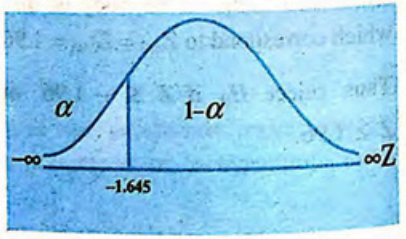
iv. Computation:

Given n = 25, $\bar{X} = 250, \sigma = 25$

$$Z = \frac{250-285}{25/\sqrt{25}} = \frac{-35}{25/5} = \frac{-35}{5} = -7$$

v. Critical region:

Given $\alpha = 0.05$ since we have one tailed test, so 0.5 – 0.05 = 0.4500. Make inverse use of area table we have $Z_\alpha = Z_{0.05} = 1.645$ i.e. reject H_0 if Z calculated is less than or equal to -1.645.



vi. Conclusion:

Since our calculated value of Z lies in the rejection region, therefore, we reject H_0 and accept H_1.

Example 7.3

Past records show that the average score of students in statistics is 57 with standard deviation 10. A new method of teaching is employed and a random sample of 70 students is selected. The sample average is 60. Can we conclude on the basis of these results, at 5% level of significance, that the average score has increased?

Solution:

i. $H_0: \mu = 57$

$H_1: \mu > 57$

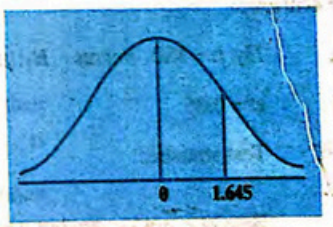
ii. $\alpha = 5\% = 0.05$

iii. Test statistic: since σ is known, therefore, test statistic for μ in this case is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

iv. Calculation: Given that $n = 70$, $\bar{X} = 60$, $\sigma = 10$, therefore,

$$Z = \frac{60-57}{10/\sqrt{70}} = \frac{(3)\sqrt{70}}{10} = 2.51$$



v. Critical region:

As we see in H_1, we need to use one-sided test to the right. Here $\alpha = 0.05$ (from the area table of S.N.D, we have

$Z_\alpha = Z_{0.05} = 1.645$).

Hence reject H_0 if $Z \geq 1.645$.

vi. Conclusion:

We see that $Z = 2.51$ lies in the rejection region, so we reject H_0 and conclude that the average score has increased.

7.2.1 Procedure for hypothesis testing about μ when σ is unknown [large sample]

When σ is unknown but $n \geq 30$, then central limit theorem allows us to consider sampling distribution of $\bar{x}$ as approximately normal with mean μ and S.E $= \frac{S}{\sqrt{n}}$. Hence the test statistic is $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ i.e. only σ is replaced by S. All other steps of hypothesis testing procedure are exactly the same as when σ^2 is known.

Example 7.4

The daily yield for a local chemical plant has average 880 tons for the last several years. The quality control manager would like to know whether this average has changed in recent months. She randomly selects 50 days from the computer database and computes the average and standard deviation of the $n = 50$ yields as $\bar{X} = 871$ tons and $S = 21$ tons respectively. Test the appropriate hypothesis using $\alpha = 0.05$

Solution:

i. $H_0: \mu = 880$ versus $H_1: \mu \neq 880$

ii. $\alpha = 0.05$

iii. Test statistic:

Here σ is not known but $n = 50$ is large, so we use $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$

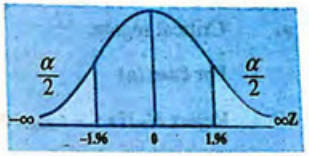
as test statistic for μ.

iv. Calculation:

$$Z = \frac{871 - 880}{21/\sqrt{50}} = -3.03$$

v. Rejection region:

For two-tailed test we use $\alpha/2 = 0.025$ so $0.5 - 0.025 = 0.4750$ which corresponds to $Z_{\alpha/2} = 1.96$ in the area table of S.N.D. We will reject if $Z \leq -1.96$ or $Z \geq 1.96$



vi. Conclusion:

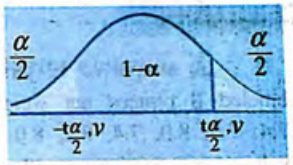
Since $Z = -3.03$ falls in the rejection region, therefore, we reject H_0.

7.2.2 Procedure for hypothesis testing about μ when σ is unknown [small sample]

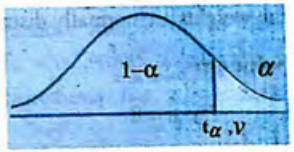
i. a) $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$

b) $H_0: \mu \leq \mu_0$ versus $H_1: \mu > \mu_0$

c) $H_0: \mu \geq \mu_0$ versus $H_1: \mu < \mu_0$

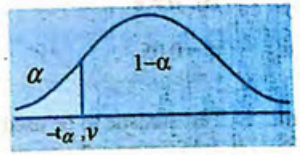


ii. Choose $\alpha = 0.01$ or 0.05 etc.



iii. Test statistic:

When σ is unknown and $n < 30$, then test statistic to be used for μ is $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ which has t-distribution with $\nu = n - 1$ degrees of freedom,



where as

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

iv. Calculation of t-value.

v. Critical region:

For case (a)

Reject H_0 if $t \le -t_{\alpha/2,(v)}$ or $t \ge t_{\alpha/2,(v)}$

For case (b)

Reject H_0 if $t \ge t_{\alpha,(v)}$

For case (c)

Reject H_0 if $t \le -t_{\alpha,(v)}$

Where t is computed value of t-statistic and $t_{\alpha/2,(v)}$ or $t_{\alpha,(v)}$ is table value.

vi. Conclusion:

Example 7.5

A sample of 12 jars of butter was taken form a lot, each jar being labeled 8 ounces net weight. The individual weights in ounces are 7.3, 7.4, 7.5, 8.0, 7.4, 8.2, 8.0, 7.6, 7.6, 7.5, 7.5, and 7.7. Test whether these values are consistent with a population mean of 8 ounces. Assume that the weights are normally distributed.

Solution:

i. H_0: $\mu = 8$

H_1: $\mu \neq 8$

ii. $\alpha = 0.05$

iii. Test statistic to be used in this case is $t=\frac{\bar{X}-\mu}{s/\sqrt{n}}$ with $v = n - 1$ d.f

iv. Calculation:

Given $n = 12$, $\Sigma x = 91.7$, $\Sigma x^2 = 701.61$, $\bar{X}=\frac{\Sigma x}{n}=\frac{91.7}{12}=7.64$

$$s=\sqrt{\frac{1}{n-1}\left[\Sigma x^2-\frac{(\Sigma x)^2}{n}\right]}=\sqrt{\frac{1}{12-1}\left[701.61-\frac{(91.7)^2}{12}\right]}$$

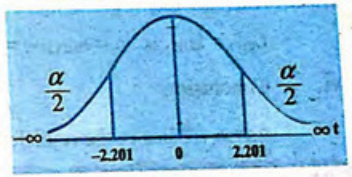
$$=\sqrt{\frac{1}{11}[0.8692]}=0.28$$

Hence $t=\frac{7.64-8}{0.28/\sqrt{12}}=\frac{(-0.36)\sqrt{12}}{0.28}=-4.45$

v. Critical region:

Reject H_0 if $t \le -t_{\alpha/2,(v)}$ or $t \ge t_{\alpha/2,(v)}$ whereas from t-table 6.1, we have

$t_{\alpha/2,(v)} = t_{0.05,(12-1)} = t_{0.025,(11)} = 2.201$



vi. Conclusion:

As $t = -4.45$ falls in the rejection region, therefore, we reject H_0.

Example 7.6

A manufacturing company making automobile tire claims that the average life of its product is 35000 miles. A random sample of 16 tires was selected and it was found that the mean life was 34000 miles with a standard deviation s = 2000 miles. Test the hypothesis H_0: $\mu = 35000$ against the alternative H_1: $\mu < 35000$ at $\alpha = 0.05$.

Solution:

i. H_0: $\mu = 35000$

H_1: $\mu < 35000$

ii. $\alpha = 0.05$

iii. Test statistic:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}, \quad v = n - 1 \text{ d.f}$$

iv. Calculation:

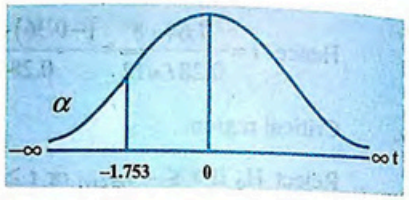
$$t = \frac{34000 - 35000}{2000/\sqrt{16}} = \frac{-1000}{2000/4} = -2$$

v. Critical region:

Reject H_0 if $t \leq -t_{\alpha,(v)}$

Whereas as from t-table

$t_{\alpha,(v)} = t_{0.05,(16-1)} = t_{0.05,(15)} = 1.753$



vi. Conclusion

As $t = -2$ lies in the rejection region, so we reject H_0

7.2.3 Procedure for testing hypothesis about difference between two populations means $(\mu_1 - \mu_2)$ when σ_1^2 and σ_2^2 are known

The samples are randomly and independently selected from the two normal populations. The formal testing procedure is explained below:

i. (a) $H_0: \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$

$H_1: \mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$

(b) $H_0: \mu_1 \leq \mu_2$ or $\mu_1 - \mu_2 \leq 0$

$H_1: \mu_1 > \mu_2$ or $\mu_1 - \mu_2 > 0$

(c) $H_0: \mu_1 \geq \mu_2$ or $\mu_1 - \mu_2 \geq 0$

$H_1: \mu_1 < \mu_2$ or $\mu_1 - \mu_2 < 0$

ii. Level of significance is decided as 0.01 or 0.05

iii. Test statistic to be used in this case is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \text{ under } H_0.$$

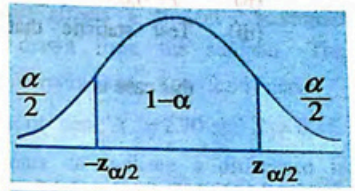
iv. Z-statistic is computed from the given data.

v. Critical region:

For case (a) in H_1

Reject H_0 if

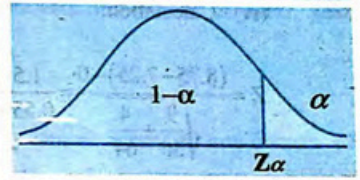
$Z \leq -Z_{\alpha/2}$ or $Z \geq Z_{\alpha/2}$



For case (b)

Reject H_0 if

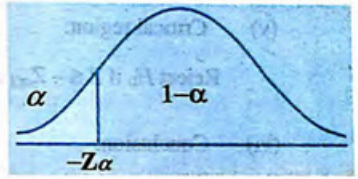
$Z \geq Z_\alpha$



For case (c)

Reject H_0 if

$Z \leq -Z_\alpha$



vi. Conclusion:

H_0 is rejected when computed value of Z falls in the rejection region.

Example 7.7

To see the effects of a certain sleeping pills on male and female, two independent samples were taken and the following data were recorded.

	male	female
sample size	$n_1 = 36$	$n_2 = 64$
sample mean	$\bar{x}_1 = 8.75$	$\bar{x}_2 = 7.25$
population variance	$\sigma_1^2 = 9$	$\sigma_2^2 = 4$

Test H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$ at 5% significance level.

Solution:

(i) H_0: $\mu_1 = \mu_2$ means that $\mu_1 - \mu_2 = 0$

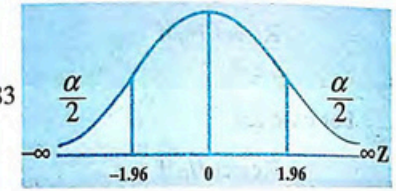
H_1: $\mu_1 \neq \mu_2$

(ii) $\alpha = 0.05$

(iii) Test statistic that summarizes the sample information in this case is $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(iv) Calculation:

$$Z = \frac{(8.75 - 7.25) - 0}{\sqrt{\frac{9}{36} + \frac{4}{64}}} = \frac{1.5}{0.559} = 2.683$$



(v) Critical region:

Reject H_0 if $Z \leq -Z_{\alpha/2}$ or $Z \geq Z_{\alpha/2}$ whereas $Z_{\alpha/2} = Z_{0.025} = 1.96$

(vi) Conclusion:

As Z = 2.683 falls in the rejection region, therefore, we reject H_0 and conclude that effect of sleeping pill on male and female is different.

7.2.4 Procedure for testing hypothesis about $(\mu_1 - \mu_2)$ when σ_1^2 and σ_2^2 are unknown but sample sizes are large.

When σ_1^2 and σ_2^2 are unknown, then for n_1, $n_2 \geq 30$, they are estimated by S_1^2 and S_2^2 respectively. The test statistic Z in this case is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1) \text{ approximately.}$$

Rest of the procedure for testing H_0 is same.

Example 7.8

To determine whether car ownership affects a student's academic achievement, two random samples were drawn from the students. The grade point average for the n_1 = 100 car owners had $\bar{X}_1$ = 2.54 and S_1^2 = 0.40 while for the n_2 = 100 non-owners of cars, $\bar{X}_2$ = 2.70 and S_2^2 = 0.36.

Do the data present sufficient evidence to indicate a difference in the mean achievements between car owners and non-owners of cars? Test using $\alpha = 0.05$.

Solution:

Let μ_1 denote mean of car owners and μ_2 denote mean of non-owners of cars, then

i. H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

ii. $\alpha = 0.05$

iii. Test statistics:

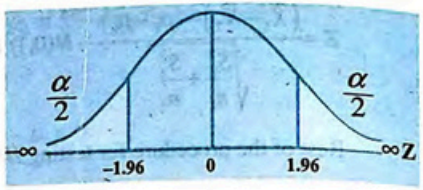
Since σ_1^2, σ_2^2 are not given but n_1, n_2 are large, so Z-statistic will be of the form $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

iv. Computations: Substituting values into the formula we get

$$Z = \frac{(2.54 - 2.70) - 0}{\sqrt{\frac{0.40}{100} + \frac{0.36}{100}}} = -1.84$$

v. Critical region:

Reject H_0 if $Z \geq Z_{\alpha/2}$ or $Z \leq -Z_{\alpha/2}$ where $Z_{\alpha/2} = 1.96$



vi. Conclusion:

We see that $Z = -1.84$ is lying in the acceptance region. Hence H_0 cannot be rejected.

7.2.5 Procedure for testing hypothesis about $(\mu_1 - \mu_2)$ when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but unknown [small samples]

When population variances are unknown and samples are of small size then the following assumption are made to use a two-sample t-test.

(i) The sampled populations are normal.

(ii) σ_1^2 and σ_2^2 are assumed to be equal but unknown

(iii) The samples are drawn randomly.

Procedure for testing H_0 is given below:

i. (a) $H_0: \mu_1 - \mu_2 = 0$ versus $H_1: \mu_1 - \mu_2 \neq 0$

(b) $H_0: \mu_1 - \mu_2 \geq 0$ versus $H_1: \mu_1 - \mu_2 < 0$

(c) $H_0: \mu_1 - \mu_2 \leq 0$ versus $H_1: \mu_1 - \mu_2 > 0$

ii. Choose $\alpha = 0.01$ or 0.05 etc.

iii. Test statistic to be used in this case is

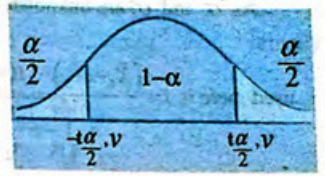
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$,Which has t-distribution with $\nu = n_1 + n_2 - 2$ d.f.

iv. Computation of t-statistic value.

v. Critical region:

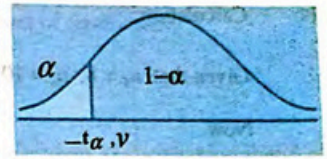
For case (a)

Reject H_0 if $t_c \leq -t_{\alpha/2,\nu}$ or $t \geq t_{\alpha/2,\nu}$



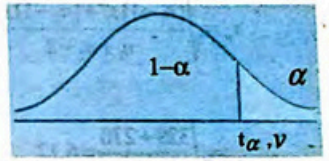
For case (b)

Reject H_0 if $t_c \leq -t_{\alpha,\nu}$



For case (c)

Reject H_0 if $t_c \geq t_{\alpha,\nu}$



vi. Conclusion.

Example 7.9

An examination was given to two classes of 8 and 10 students respectively. In the first class mean grade was 95 with a standard deviation of 6.8556, while in the second class, the mean grade was 97 with a standard deviation of 5.4772. It is assumed that the two classes of students are normally distributed having identical variances. Is there a significance difference between the mean grades? Test at $\alpha = 0.01$.

Solution:

i. $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$

ii. $\alpha = 0.01$

iii. Test statistic:

As σ_1^2 and σ_2^2 are unknown and n_1, n_2 are small ,therefore, test statistic to be used here is $t=\dfrac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ with $v=n_1+n_2-2$ d.f

iv. Calculation:

Given that $n_1=8,\ n_2=10,\ \bar{X}_1=95,\ \bar{X}_2=97,\ s_1^2=47,\ s_2^2=30$

Now

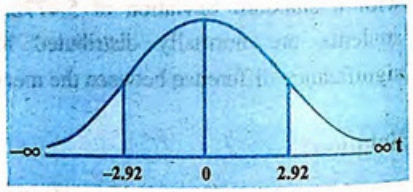
$$s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}=\sqrt{\frac{(8-1)47+(10-1)30}{8+10-2}}$$

$$=\sqrt{\frac{329+270}{16}}=6.12$$

Hence $t=\dfrac{(95-97)-0}{6.12\sqrt{\dfrac{1}{8}+\dfrac{1}{10}}}=\dfrac{-2}{2.930}=-0.689$

v. Critical region:

Reject H_0 if $t \le -t_{\alpha/2,v}$ or $t \ge t_{\alpha/2,v}$, whereas $t_{\alpha/2,v}$ = $t_{0.01/2,\ 8+10-2}=t_{0.005,\ (16)}=2.92$



vi. Conclusion:

Our calculated value of t falls in the acceptance region, so we accept H_0.

Example 7.10

The following two samples are drawn from the normally distributed population with identical but unknown variances.

Sample 1: 70 68 63 60 59 57 53 51 50 49 46 45

Sample 2: 75 72 70 50 48 43 52 50 46 45

Test the equality of means at $\alpha = 0.05$ level of significance.

Solution:

i. $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

ii. $\alpha = 0.05$

iii t-test statistic to be used here is

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$$ with $v=n_1+n_2-2$ degrees of freedom

(iv) Calculation:

For sample-1:

$n_1=12,\ \sum x_1=671,\ \sum x_1^2=38275,\ \bar{X}_1=55.92$

$$s_1^2=\frac{1}{n_1-1}\left[\sum x_1^2-\frac{(\sum x_1)^2}{n_1}\right]$$

or $(n_1-1)s_1^2=\sum x_1^2-\dfrac{(\sum x_1)^2}{n_1}=38275-\dfrac{(671)^2}{12}=754.9167$

For sample-2:

$n_2=10,\ \sum x_2=551,\ \sum x_2^2=31707,\ \bar{X}_2=55.1$

$$s_2^2=\frac{1}{n_2-1}\left[\sum x_2^2-\frac{(\sum x_2)^2}{n_2}\right]$$

$$\text{Or } (n_2-1)s_2^2=\sum x_2^2-\frac{(\sum x_2)^2}{n_2}=31707-\frac{(551)^2}{10}=1346.9$$

Now

$$s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}=\sqrt{\frac{754.9167+1346.9}{12+10-2}}$$

$$=\sqrt{\frac{2101.8167}{20}}=10.25$$

$$\text{Hence } t=\frac{(55.92-55.1)-0}{10.25\sqrt{\frac{1}{12}+\frac{1}{10}}}=\frac{0.82}{4.3888}=0.187$$

v. Critical region:

We will reject H_0 if $t < t_{\alpha/2,\nu}$ of $t > t_{\alpha/2,\nu}$ where as

$t_{\alpha/2,\nu} = t_{0.05/2,(12+10-2)} = t_{0.025,(20)} = 2.086$

vi. Conclusion:

As our computed value of t lies in the acceptance region, therefore, we accept H_0.

7.2.6 Procedure of testing hypothesis about population pro ortion p [large sample]

For large n, ($n \geq 30$), and the hypothesis testing procedure is out lined as below:

i. (a) $H_0: p = p_0$ versus $H_1: p \neq p_0$

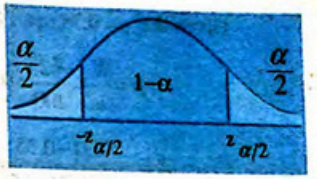
(b) $H_0: p \leq p_0$ versus $H_1: p > p_0$

(c) $H_0: p \geq p_0$ versus $H_1: p < p_0$

ii. Choose $\alpha = 0.01$ or 0.05 etc.

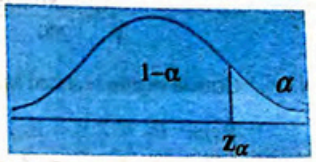
iii. Test statistic to be used in this case is

$$Z=\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \text{ under } H_0 \text{ when } n \geq 30$$



iv. Calculation:

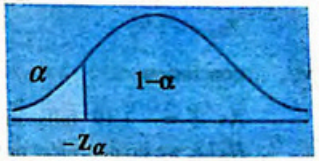
v. Critical region according to H_1 is stated as



(a) Reject H_0 if $Z \leq -Z_{\alpha/2}$ or $Z \geq Z_{\alpha/2}$

(b) Reject H_0 if $Z > Z_\alpha$

(c) Reject H_0 if $Z < -Z_\alpha$



vi. Conclusion.

Example 7.11

A producer of orange juice claims that 35 percent of all orange juice drinkers prefer its product. To test the claim, a random sample of 200 orange juice drinkers was taken at random and it was found that only 62 of them preferred the producer's brand. Test the producer's claim at five percent level of significance.

Solution:

i. $H_0: p = 35\% = 0.35$

$H_1: p \neq 0.35$

ii. $\alpha = 0.05$

iii. Test statistic to be used here is $Z=\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$

iv. Calculation:

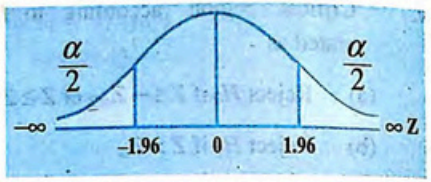
Given that p= 0.35, q= 1 – p = 1 – 0.35 = 0.65, n = 200, X = 62,

$$\therefore \hat{p}=\frac{X}{n}=\frac{62}{200}=0.31,$$

$$Z=\frac{0.31-0.35}{\sqrt{\frac{(0.35)(0.65)}{200}}}=-1.19$$

v. Critical region is stated as:

Reject H_0 if $Z \leq -Z_{\alpha/2}$ or $Z \geq Z_{\alpha/2}$ where $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$



vi. conclusion:

As Z =– 1.19 falls in the acceptance region, therefore, we accept H_0 i.e. we do not reject producer's claim is correct.

7.2.7 Procedure for testing hypothesis about difference of two population proportions (p_1–p_2) [large sample]

The testing of hypothesis procedure in this case is given below.

i. (a) $H_0: p_1 - p_2 = 0$ (b) $H_0: p_1 - p_2 \leq 0$ (c) $H_0: p_1 - p_2 \geq 0$

$H_1: p_1 - p_2 \neq 0$ $H_1: p_1 - p_2 > 0$ $H_1: p_1 - p_2 < 0$

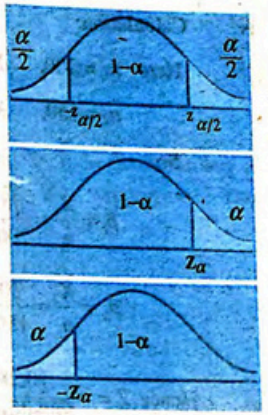
ii. Choose an appropriate level of significance.

iii. Z-test is used where as

$$Z=\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{p_2\hat{q}_2}{n_2}}} \sim N(0,1) \text{ under } H_0.$$

iv Calculation:

v. Critical region is constructed according to H_1 as;



Case (a) Reject H_0 if $Z_c \leq -Z_{\alpha/2}$ or $Z_c \geq Z_{\alpha/2}$ i.e.

Case (b) Reject H_0 if $Z_c \geq Z_\alpha$ i.e.

Case (c) Reject H_0 if $Z_c \leq -Z_\alpha$ i.e.

v. Conclusion.

Example7.12

A soap-manufacturing factory produces two brands of soap. A sample survey was conducted and it was found that 56 users out of 200 preferred brand "A" and that 30 users out of 150 preferred brand "B". Test the hypothesis that sale of brand A is at least 10% greater than brand B at 5% level of significance.

Solution:

Let p_1 denote the proportion of brand-A users and p_2 denote the proportion of brand-B users.

i. $H_0: p_1 - p_2 \geq 0.10$

$H_1: p_1 - p_2 < 0.10$

ii. $\alpha = 0.05$

iii. Test statistic:

From H_0 we see that $p_1 \neq p_2$ and they are unknown therefore they are estimated by $\hat{p}_1$ and $\hat{p}_2$ respectively. The test statistic to be used is;

$$Z=\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}} \sim N(0,1) \text{ under } H_0.$$

iv. Calculation:

Here $n_1 = 200, \quad X_1 = 56$

$n_2 = 150, \quad X_2 = 30$

So

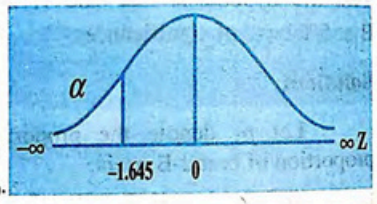
$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{56}{200} = 0.28 \ , \ \hat{q}_1 = 1 - \hat{p}_1 = 1 - 0.28 = 0.72$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{30}{150} = 0.2 \ , \ \hat{q}_2 = 1 - \hat{p}_2 = 1 - 0.2 = 0.8$$

Hence $$Z = \frac{(0.28 - 0.2) - 0.10}{\sqrt{\frac{(0.28)(0.72)}{200} + \frac{(0.2)(0.8)}{150}}} = \frac{-0.02}{0.0455} = -0.44$$

v. Critical region:

Reject H_0 if Z-calculated value is less than $-Z_\alpha = -1.645$



vi. Decision:

As computed value of Z falls in the acceptance region, so we accept H_0.

Example 7.13

The records of a hospital show that 52 men in a sample of 1000 men versus 23 women in a sample of 1000 women were admitted because of heart disease. Do these data present sufficient evidence to indicate a higher rate of heart disease among men admitted to the hospital? Use $\alpha = 0.05$.

Solution:

Let p_1 denote the proportion of men with heart disease and p_2 denote the proportion of women with heart disease.

i $H_0: p_1 \le p_2$ or $p_1 - p_2 \le 0$

$H_1: p_1 > p_2$ or $p_1 - p_2 > 0$

ii $\alpha = 0.05$

iii Test statistic:

From H_0 we see that $p_1 = p_2$ but unknown. So it is better to use a pooled estimate for the unknown value of the proportion. The estimated test-statistic to be employed here is $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_c \hat{q}_c (\frac{1}{n_1} + \frac{1}{n_2})}}$

iv. Calculation:

Given $n_1 = 1000, X_1 = 52, \ \hat{p}_1 = \frac{X_1}{n_1} = \frac{52}{1000} = 0.052,$

$n_2 = 1000, X_2 = 23, \ \hat{p}_2 = \frac{X_2}{n_2} = \frac{23}{1000} = 0.023$

Now $$\hat{p}_c = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{52 + 23}{1000 + 1000} = \frac{75}{2000} = 0.0375$$

And $\hat{q}_c = 1 - \hat{p}_c = 1 - 0.0375 = 0.9625$

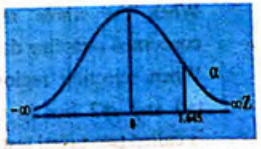
Hence $$Z = \frac{(0.052 - 0.023) - 0}{\sqrt{(0.0375)(0.9625)\left(\frac{1}{1000} + \frac{1}{1000}\right)}} = 3.41$$

v. Critical region:

Reject H_0 if $Z > Z_\alpha$, whereas $Z_\alpha = Z_{0.05} = 1.645$

vi. Conclusion:

$Z = 3.41$ falls in the rejection region, so we reject H_0. This observed data indicates that the percentage of men entering the hospital because of heart disease is higher than that of women.



Key points

- Hypothesis is an unproved claim or assertion or assumption which acts as a starting point in a research irrespective of its probable truthfulness or falsity.
- A statistical hypothesis is a testable claim about one or more parameter(s) of empirical distributions.
- A statistical method that uses sample data to accept or reject a hypothesis about a parameter is called hypothesis testing.
- Null hypothesis is a claim about parameter. It is tested for possible rejection under the assumption it is true. It is denoted by H_0
- Alternative hypothesis is also a claim about parameter which is accepted if H_0 is rejected. It is tested for possible acceptance under the assumption it is false. It is denoted by H_1.
- When all parameters of a distribution are well specified, it is called simple hypothesis.
- When all parameters of a distribution are not well specified, it is called composite hypothesis.
- "Reject a null hypothesis when it is true" is called Type-I error.
- "Accept a null hypothesis when it is false" is called Type-II error.
- The probability of Type-I error which we are ready to tolerate in making decision about H_0 is called significance level
- A function of sample data is called statistic. When a statistic is used to test a hypothesis it is called test statistic or decision rule.
- The group of values which would lead us to acceptance of H_0 or the part where $(1 - \alpha)$ lies is called acceptance region (AR) for the test.
- The group of values which would lead us to rejection of H_0 or the part where (α) lies is called rejection region (RR) or critical region for the test.
- The values which separate the (AR) and (RR) are called critical values.
- When the whole rejection region of size α lies on one tail of the concerned sampling distribution, it is called one-tailed test.
- When rejection region of size α is equally divided, $\alpha/2$ lies in the right tail and $\alpha/2$ lies in the left tail of the concerned sampling distribution, it is called two-tailed test.

Exercise

7.1 Write "T" for true and "F" for false in the following statement.

i. A statistical hypothesis is a statement about the value of a statistic.
ii. The null hypothesis is framed for possible rejection.
iii. In statistical inference "accept H_0" means that there is insufficient information to reject H_0.
iv. An upper-tailed test occurs when H_0: $\mu \geq \mu_0$ and H_1: $\mu < \mu_0$.
v. If a null hypothesis H_0: $\mu = 50$ is rejected at 1% level of significance, it will also be rejected a 5% level of significance.
vi. When a true null hypothesis has been rejected, we say that Type-I error might have been committed.
vii. α and β have an inverse relationship.
viii. H_0: $\mu < 5$ is a simple null hypothesis.
ix. When the sample size n increases the probability of rejecting a true hypothesis decreases.
x. Confidence interval estimate of a parameter is the same thing as testing of hypothesis about the population parameter.

7.2 Fill in the suitable word in the blanks.

i. A hypothesis is an _____ about the parameter of a population.
ii. The hypothesis which is under test for possible rejection is called _____ hypothesis.
iii. A hypothesis contrary to null hypothesis is known as _____.
iv. The hypothesis H_0: $\mu > \mu_0$ is a _____ hypothesis.
v. Type _____ error is more severe than type _____ error.
vi. Probability of Type-I error is called _____.
vii. Level of significance lies between _____ and _____.
viii. Critical region is also known as _____.
ix. A statistical test is a _____ to decide about H_0.
x. The number of independent values in a set of values is known as - _____.

7.3 Choose the correct answer.

i. A hypothesis under test is:
(a) Simple hypothesis (b) composite hypothesis
(c) Null hypothesis (d) alternative hypothesis

ii. Whether a test is one-sided or two-sided depends on:

(a) Alternative hypothesis (b)Composite hypothesis

(c) null hypothesis (d) Simple hypothesis

iii. A wrong decision about H_0 leads to:

(a) One kind of error (b) Two kind of error

(c) Three kind of error (d) Four kind of error.

iv. Level of significance is the probability of:

(a) Type-I error (b) Type-II error

(c) Not committing error (d) Any of the above

v. Degrees of freedom is related to:

(a) Number of observations in a set

(b) Hypothesis under test

(c) Number of independent observation in set

(d) confidence interval

vi. As compared to normal distribution, the t-distribution is:

(a) Flatter (b) More peaked

(c) Symmetric (d) Negatively skewed.

vii. Student's t-test is applicable in case of:

(a) Small samples

(b) Large samples

(c) For samples of size between 5 and 29.

(d) biased samples

viii. To test H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$ when the population S.D is known, the appropriate test is:

(a) t-test (b) Z-test

(c) Chi-square (d) F-test.

ix. Testing hypothesis H_0: $\mu = 10$ vs H_1: $\mu > 10$ leads to:

(a) One sided left tailed test (b) one sided right tailed test

(c) Two-tailed test (d) all of the above

x. Range of statistic-t is:

(a) -1 to $+1$, (b) $-\infty$ to ∞

(c) 0 to ∞ (d) 0 to 1

7.4 Define hypothesis, statistical hypothesis and testing of hypothesis?

7.5 Explain in your own words (i) Hypothesis testing (ii) Estimation. What are the principle steps involved in hypothesis testing procedure?

7.6 Differentiate between:

(i) Null and alternative hypothesis.

(ii) Simple and composite hypothesis.

(iii) Type-I and Type-II Errors.

(iv) Acceptance and rejection region.

7.7 Explain the following terms:

(a) Level of significance.

(b) Test statistic and test of significance.

(c) Small sample and large sample.

7.8 What do you mean by one-sided test and two sided test? Explain your idea with diagrams?

7.9 Describe the general procedure (steps) for testing hypothesis about mean of a normal population when population standard deviation is known and sample size is large.

7.10 A random sample of size n = 900 plants and its mean is computed which is equal to 34 cm. Can it be reasonably regarded as a random sample from a large population with mean 32 cm and standard deviation 23 cm. Testing at $\alpha = 0.05$.

7.11 Let $\bar{X} = 15$ be the mean of a random sample of 64 observations drawn from a normal population whose variance is 100. Test H_0: $\mu = 12$ versus H_1: $\mu > 12$ at 5% level of significance.

7.12 If $\bar{X} = 42.6$ is the mean of a random sample of size 36 taken from a normal population with a known standard deviation $\sigma = 5$. Test the null hypothesis $\mu = 45$ against the alternative $\mu < 45$ using $\alpha = 0.05$.

7.13 Test the hypothesis that population mean is 150 when $n = 196$, $\bar{X} = 160$, $S = 60$ using $\alpha = 0.05$.

7.14 A random sample of 49 college students showed an average IQ of $\bar{X} = 120.67$ and $S = 8.44$. Test the hypothesis that the average IQ of the college students is equal to 123 against the alternative that it is less. Test at 5% level of significance.

7.15 Explain hypothesis testing procedure about mean of a normal population when σ is unknown and sample size is small.

7.16 A random sample of size 10 is drawn from a normal population gives $\bar{X} = 20$ and $s^2 = 16$. Test the hypothesis H_0: $\mu = 19.6$ against H_1: $\mu > 19.6$ at $\alpha = 0.05$.

7.17 The nine items of a sample had the following values 45, 47, 50, 52, 48, 47, 49, 53, 51. Does the mean of the nine items differ significantly from an assumed population mean of 47.5? Test at 5% level of significance.

7.18 Describe the procedure for testing hypothesis about the equality of means of two normal populations when population standard deviations are known and sample sizes are large.

7.19 A random sample of size $n_1 = 25$ taken from a normal population with a standard deviation $\sigma_1 = 5.2$ has a mean $\bar{X}_1 = 81$. A second random sample of size $n_2 = 36$ taken from a different normal population with a standard deviation $\sigma_2 = 3.4$ has a mean $\bar{X}_2 = 76$. Test the hypothesis at the 0.05 level of significance that $\mu_1 - \mu_2 = 0$ against the alternative $\mu_1 - \mu_2 \neq 0$.

7.20 Independent random samples are drawn from two quantitative populations, 1 and 2 respectively. The sample data summary is shown here:

	sample-1	sample-2
sample size	36	45
sample mean	1.24	1.31
sample variance	0.0560	0.0540

Do the data present sufficient evidence to indicate that the mean for population-1 is smaller than the mean for population-2? Use $\alpha = 0.05$

7.21 Write the steps used in hypothesis testing about equality of means of two normal populations when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but unknown and sample sizes n_1, n_2 are small.

7.22 Test the hypothesis that the mean number of kilometers per liter is the same for foreign and domestic auto-mobiles on the basis of the following summary of sample data:

	n	$\bar{x}$	s^2
Foreign automobiles	8	36.5	5.29
Domestic automobiles	10	32.4	7.84

Test at 5% level of significance.

7.23 Given the following sample observations

X_1	17	27	18	25	27	29	27	23	17
X_2	16	16	20	16	21	17	15	20	

Examine the significance of difference between the two population means at $\alpha = 0.05$

7.24 Explain the general procedure for testing hypothesis about population proportion P for a large sample.

7.25 A random sample of 120 observations was selected from a binomial population, and 72 successes were observed. Do the data provide sufficient evidence to indicate that p is greater than 0.5? Use $\alpha = 0.05$.

7.26 Blood donation society of a college claimed that 5% of the students in our college make blood donation during a given year. If in a random sample 10 of 250 students have given blood during the past year. Test H_0: $p = 0.05$ against H_1: $p \neq 0.05$ with $\alpha = 0.05$.

7.27 An electric company claimed that at least 85% of the parts which it supplied confirmed to specifications. A sample of 400 parts was tested and 75 did not meet specifications. Can we accept the company's claim at 1% level of significance?

7.28 Describe the steps used in hypothesis testing about difference of two population proportions in case of large samples.

7.29 Test the hypothesis that proportions of men and women favoring a political candidate are different on the basis of a sample survey in which 225 of 500 men and 275 of 500 women favor the candidate. Test at 5% level of significance.

7.30 In a random table of 600 persons from a certain large city 450 are found to be smokers. In another sample of 900 persons from another large city 450 are smokers. Test at $\alpha = 0.01$ that the two cities are significantly different with respect to the prevalence of smoking.

Unit - 8

Association of Attributes

After studying this unit, the students will be able to

- Recall variable and attribute.
- Recognize the notation and terminology to represent the presence and absence of attributes.
- Describe class and class frequency.
- Recognize the categorical data of two attributes.
- Explain independence of two attributes.
- Know the criterion of independence of two attributes.
- Discuss the association of two attributes; positive association, negative association, complete association and complete disassociation.
- Define Yule's coefficient of association.
- Find the coefficient of association and interpret its result.
- Define contingency table.
- Test whether two attributes in a given contingency table are statistically independent or not.
- Describe Pearson's coefficient of mean square contingencies.
- Calculate Pearson's co efficient of mean square contingency for a given contingency table and find its maximum value.
- Describe and apply Yule's correction for continuity to test the statistical independence of two given attributes.

8.1 Introduction to attributes

In our daily life, we study the characteristics like gender, health, satisfaction, religion, colour etc. that cannot be measured and expressed quantitatively but instead of its qualitative or descriptive nature, only their presence or absence can be noticed. These descriptive or qualitative characteristics are called attributes. Attributes cannot be measured accurately but they can be divided into classes and their numbers in each class can be counted e.g. the above characteristics can be classified as male or female, healthy or unhealthy, satisfied or unsatisfied, Muslim or non-Muslim, white or black etc.

8.1.1 Notation and terminology for attributes

- **Symbols**

For the sake of convenience capital English letters A, B, C... are used to denote presence of attributes and the Greek letters α, β, γ... denote the absence of these attributes respectively. For example if A represents the class "Muslim", then α will represent the class "non-Muslim". Similarly, if we are studying two or more attributes, their combination can be represented by combining the letters representing the attributes. For example, if A represent "blindness" and B "deafness", then AB will represent the "blindness and deafness". Similarly, if A represents the "Muslim" and B "male", then the following four classes will be formed for the presence or absence of these attributes

i) Muslim and male = AB

ii) Muslim and female = Aβ

iii) Non-Muslim and male = αB

iv) Non-Muslim an female = $\alpha\beta$

- **Positive and negative classes**

The classes A, B, AB are called positive classes because they contain the presence of attributes. The classes $\alpha, \beta, \alpha\beta$ are called negative classes because they contain absence of the attributes. The classes Aβ, αB contain both presence and absence of the attributes, hence are called mixed classes.

- **Class frequency**

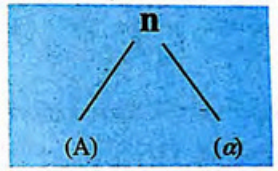
The number of observations falling in a class is called frequency of the class. For attributes class frequencies are denoted by enclosing the class symbols in parentheses.

Thus (A) denotes frequency of class/attribute A. (B) denotes frequency of attributes B. Similarly, (AB) denotes the number of individuals / objects possessing both attributes A and B. (ABC) is frequency for class ABC and so on.

8.1.2 Classification of attributes

- **Dichotomy or one-way classification**

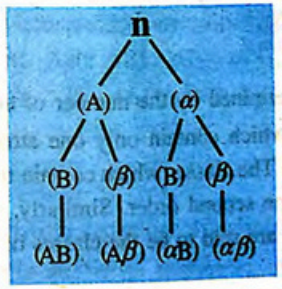
When a single attribute is under study, then simple process of classifying the whole data in two groups is called dichotomy which means classifying into two.



Since only one attribute is involved, so the division of data is called one-way classification.

- **Two-way classification**

When two attributes A and B are under study then whole data are classified into four classes or groups as follows:



Since two attributes are involved, the division of the sample is called two-way classification.

• 2 × 2 Contingency table

Classification of data about two attributes each having two classes or categories can be shown in tabular form as given below.

(2 × 2) Contingency table

Attributes	B	β	Row total
A	(AB)	(Aβ)	(A)
α	(αB)	($\alpha\beta$)	(α)
Column total	(B)	(β)	n

Since the table contains 2 rows and 2 columns is therefore called 2 × 2 contingency table.

Remember that:

$n = (A) + (\alpha)$

$n = (B) + (\beta)$

$(A) = (AB) + (A\beta)$

$(\alpha) = (\alpha B) + (\alpha\beta)$

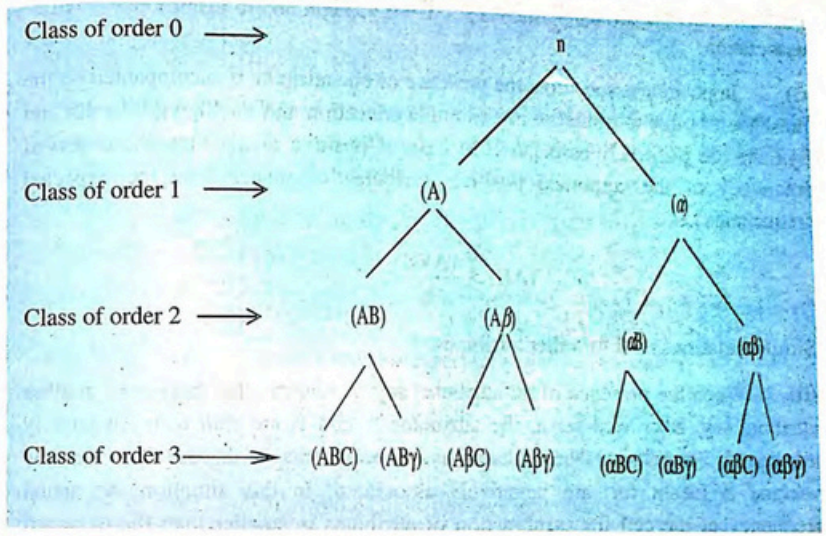
$(B) = (AB) + (\alpha B)$

$(\beta) = (A\beta) + (\alpha\beta)$

• Order of classes

The order of a class is determined by the number of attributes present in a class. For example, the classes which contain only one attribute say "A" or "B" are called classes of first order. The classes which contain two attributes say "AB" are said to be the classes of the second order. Similarly, the classes which contain three attributes say "ABC" are said to be the classes of the third order and so on.

The sample size n does not contain any attributes so is called class of order zero. Let us understand the above discussion through tree diagram as;



• Ultimate classes and ultimate class frequencies

The classes and class frequencies of the highest order are called the ultimate classes and ultimate class frequencies. For example, for one attribute ultimate classes are A, α and ultimate frequencies are (A), (α). For two attributes, ultimate classes are AB, Aβ, αB, $\alpha\beta$ and ultimate classes frequencies are (AB), (Aβ), (αB), ($\alpha\beta$). If we are considering three attributes A, B, C, then ultimate classes are ABC, ABγ, AβC, A$\beta\gamma$, αBC, αBγ, $\alpha\beta$C, $\alpha\beta\gamma$ and ultimate class frequencies are (ABC), (ABγ), (AβC), (A$\beta\gamma$), (αBC), (αBγ), ($\alpha\beta$C), ($\alpha\beta\gamma$)and so on.

8.1.3 Association of attributes

Association is a statistical technique which measures the strength and direction of relationship among qualitative variables.

• Kinds of association

There are three kinds of associations which ossibly occur between attributes namely (i) positive association, (ii) negative association and (iii) no association.

(i) In positive association, the presence of one attribute is accompanied by the presence of other attribute(s). For example education and intelligence, health and hygiene are positively associated. In case of positive association, the observed frequency of the combined positive attributes is greater than the expected frequencies i.e.

$$(AB) > \frac{(A)(B)}{n}.$$

Similar relations hold for other attributes.

(ii) When the presence of an attributes say, A ensures the absence of another attribute say, B or vice-versa, the attributes A and B are said to be negatively associated. For instance, the vaccination and occurrence of disease for which the vaccine is meant for, are negatively associated. In this situation the actual frequency of the cell for combination of attributes is smaller than the expected frequency i.e.

$$(AB) < \frac{(A)(B)}{n}.$$

(iii) If the two attributes are such that the presence or absence of one attribute has nothing to do with the presence or absence of the other, they are said to be independent. For instance, skin colour and intelligence of persons are independent attributes. If the two attributes A and B are independent, then $(AB) = \frac{(A)(B)}{n}$.

This is known as criterion of independence.

8.1.4 Methods of measures of association

There are mainly three methods of measures of association.

(i) Yule's coefficient of association.

(ii) Chi-square test for testing hypothesis about independence of attributes in contingency tables.

(iii) Coefficient of contingency for $(r \times c)$ contingency table.

• Yule's coefficient of association

Yule's coefficient of association is named after its inventor G. Undy Yule. It is a relative measure for the strength of association between two attributes say, A and B. It is defined by the formula given below.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}, \qquad -1 < Q < +1$$

The result of Q is interpreted as:

If	$Q = -1$	Attributes have perfect negative association.
	$-1 < Q < 0$	Attributes have negative association.
	$Q = 0$	Attributes have no association means they are independent.
	$0 < Q < +1$	Attributes have positive association.
	$Q = +1$	Attributes have perfect positive association.

Example 8.1

Discuss the association between two attributes say A and B when:

(i) $(AB) = 110, (\alpha B) = 96, (A\beta) = 290, (\alpha\beta) = 510$

(ii) $(A) = 245, (AB) = 147, (\alpha) = 285, (\alpha B) = 190$

Solution:

(i) Given $(AB) = 110, (\alpha B) = 96, (A\beta) = 290, (\alpha\beta) = 510$

Coefficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Q=\frac{(110)(510)-(290)(96)}{(110)(510)+(290)(96)}$$

$$Q=\frac{28260}{83940}=0.34$$

It means that there is positive association between attributes A and B.

(ii) Given (A) = 245, (AB) = 147, (α) = 285, (αB) = 190

Since all values required for the coefficient of association formula are not known, so first put these values in a (2 × 2) contingency table to find the unknown values easily by addition or subtraction as.

(2 × 2) contingency table

Attributes	B	β	Row total
A	(AB) = 147	(Aβ) = 98	(A) = 245
α	(αB) = 190	($\alpha\beta$) = 95	(α) = 285
Column total	(B) = 337	(β) = 193	n = 530

Now

$$Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$Q=\frac{(147)(95)-(98)(190)}{(147)(95)+(98)(190)}$$

$$Q=\frac{13965-18620}{13965+18620}=\frac{-4655}{32585}=-0.14$$

Thus there is negative association between A and B.

Example 8.2

The following table shows the data obtained during an epidemic of cholera.

Attributes	attacked	not attacked
inoculated	31	469
not inoculated	185	1315

Test the effectiveness of inoculation in preventing attack of cholera.

Solution:

Let us denote inoculated by A and attack by B.

The given data can be written as under

(AB) = 31, (Aβ) = 469, (αB) = 185, ($\alpha\beta$) = 1315

Putting in the coefficient of association we get

$$Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$Q=\frac{(31)(1315)-(469)(185)}{(31)(1315)+(469)(185)}=\frac{40765-86765}{40765+86765}=\frac{-46000}{127530}=-0.36$$

There is negative association between inoculation and attack of cholera disease.

Example 8.3

For admission in a medical college 1660 candidates appeared in an entry test and 422 were successful. 256 attended a coaching class and of these 150 came out successful. Estimate the utility of the coaching classes.

Solution:

Putting the given information in (2 × 2) contingency table as;

Attributes	Successful	Not successful	Total
Coached A	AB = 150	$A\beta = 106$	256
Not coached α	$\alpha B = 272$	$\alpha\beta = 1132$	1404
Total	422	1238	1660

Yule's coefficient of association is given as

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Q = \frac{(150)(1132) - (106)(272)}{(150)(1132) + (106)(272)} = \frac{140968}{198632} = 0.71$$

There is high positive association means that coaching helps in success.

8.1.5 Multi-way classification

When a population or sample is divided into many classes or categories according to an attribute is called multi-way or manifold classification. For example, a population according to "colour of eyes" can be divided as black eyes, grey eyes, green eyes etc. Population of students according to "performance" in an examination can be rated as excellent, good, average and poor. Heights of persons can be categorized as tall, medium and short. The classification of data about such attributes which have many classes can be shown by means of a contingency table.

8.2 Contingency table

A table consisting of r-rows and c-columns into which the data are classified according to two attributes is called (r × c) contingency table. For example, if an attribute "A" has $A_1, A_2, \ldots, A_r$ classes and attributes "B" has $B_1, B_2, \ldots, B_c$ classes, then the observed data by $(r \times c)$ contingency table is shown as:

$(r \times c)$ contingency table

Classes	B_1	B_2	...	B_j	...	B_c	Row total
A_1	(A_1B_1)	(A_1B_2)	...	(A_1B_j)	...	(A_1B_c)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	...	(A_2B_j)	...	(A_2B_c)	(A_2)
⋮	⋮		...	⋮	...	⋮	...
A_i	(A_iB_1)	(A_iB_2)	...	(A_iB_j)	...	(A_iB_c)	(A_i)
⋮	⋮		...	⋮	...	⋮	...
A_r	(A_rB_1)	(A_rB_2)	...	(A_rB_j)	...	(A_rB_c)	(A_r)
Column total	(B_1)	(B_2)	...	(B_j)	...	(B_c)	n

This table is an extension of (2 × 2) contingency table. The values in the cells of contingency table shown in parentheses are called cell frequencies. For each observed frequency O_f, the expected frequency E_f is computed as $E_f = \frac{R \times C}{n}$, where R is the row total and C is the column total.

For example; for observed frequency (A_1B_1) given in above table, the corresponding expected frequency will be equal to $\frac{(A_1)(B_1)}{n}$ and similarly for (A_1B_2) the expected frequency = $\frac{(A_1)(B_2)}{n}$ and so on. Chi-square test statistic denoted by χ^2 (pronounced as kai square) is used to test the hypothesis about independence of attributes in the contingency tables.

8.2.1 General procedure for testing hypothesis about independence of attributes in contingency tables

i. H_0: The attributes are independent.

H_1: The attributes are associated.

ii. $\alpha = 0.01$ or 0.05 etc.

iii. Test statistic to be used here is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \text{ with } \upsilon=(r-1)(c-1) \text{ d.f}$$

Whereas, O_{ij} denotes observed frequency

e_{ij} : expected frequency

r : number of rows

c : number of columns

iv. Calculation

v. Critical region

Reject H_0 if $\chi^2_{cal} \geq \chi^2_{tab}$, whereas χ^2 table value is obtained from χ^2 table 8.1

vi. Conclusion

Table 8.1 Critical values for the chi-square (χ^2) distribution.

df	α						
υ	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	17.275	189.675	21.920	22.618	24.725	26.757	31.264
12	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	36.741	40.113	43.194	44.140	46.963	49.645	55.476
28	37.916	41.337	44.461	45.419	48.278	50.993	56.893
29	39.087	42.557	45.722	46.693	49.588	52.336	58.302
30	40.256	43.773	46.979	47.962	50.892	53.672	59.703

Example 8.4

Consider the data given in the following contingency table

General ability \ Mathematical ability	Good B_1	Fair B_2	Poor B_3
Good	44	82	44
Fair	265	257	178
Poor	41	91	98

Discuss the association between the two attributes i.e. mathematical ability and general ability.

Solution:

i. H_0: There is no association between attributes.

H_1: There is association.

ii. We choose $\alpha = 0.05$

iii. Test statistic to be used here is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \text{ with } v=(r-1)(c-1) \text{ d.f}$$

iv. Calculation

(O_{ij})

General ability \ Mathematical ability		Good B_1	Fair B_2	Poor B_3	Total
Good	A_1	44	82	44	170
Fair	A_2	265	257	178	700
Poor	A_3	41	91	98	230
Total		350	430	320	1100

The expected frequencies are obtained by multiplying respective rows and columns totals and are divided by total sample size n as given in the table.

(e_{ij})

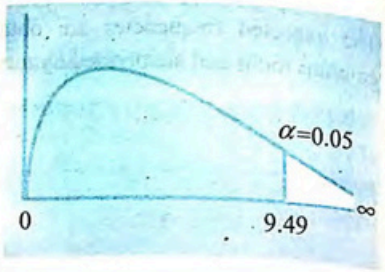
General ability \ Mathematical ability		Good B_1	Fair B_2	Poor B_3	Total
Good	A_1	$\frac{170\times350}{1100}=54.09$	$\frac{170\times430}{1100}=66.45$	$\frac{170\times320}{1100}=49.45$	169.99
Fair	A_2	$\frac{700\times350}{1100}=222.73$	$\frac{700\times430}{1100}=273.64$	$\frac{700\times320}{1100}=203.63$	700.01
Poor	A_3	$\frac{230\times350}{1100}=73.18$	$\frac{230\times430}{1100}=89.91$	$\frac{230\times320}{1100}=66.91$	230
Total		350	430	320	1100

Now χ^2 test statistic value is computed as;

O_{ij}	e_{ij}	$(O_{ij}-e_{ij})$	$(O_{ij}-e_{ij})^2$	$\frac{(O_{ij}-e_{ij})^2}{e_{ij}}$
44	54.09	−10.09	101.81	1.882
265	222.73	42.27	1786.75	8.022
41	73.18	−32.18	1035.55	14.151
82	66.45	15.55	241.80	3.639
257	273.64	−16.64	276.8896	1.012
91	89.91	1.09	1.188	0.013
44	49.45	−5.45	29.70	0.601
178	203.63	−25.64	657.41	3.228
98	66.91	31.09	966.59	14.446
1100	1100	-	-	$\chi^2 = 46.994$

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} = 46.994$$

v. Critical region



Reject H_0 if $\chi_c^2 \geq \chi_{\alpha,v}^2$ whereas from table 8.1 of χ^2 distribution, we have

$$\chi_{\alpha,v}^2 = \chi_{0.05,(3-1)(3-1)}^2 = \chi_{0.05,4}^2 = 9.49\ \chi^2$$

vi. Conclusion

Since our computed value of chi-square lies in the rejection region, therefore, we reject H_0 and conclude that there is association between mathematical ability and general ability.

8.2.2 Shortcut method of calculating χ^2 in case of (2 × 2) contingency table

When we have (2 × 2) contingency table having four cell frequencies as given below;

Attributes	B	β	Total
A	a	b	$a+b$
α	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

The value of χ^2 can be calculated directly without computing the expected frequencies by using the following formula

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad \text{with 1 d.f}$$

Example 8.5

A random sample of 100 educated and 200 uneducated people were asked about liking and disliking of football game and the following data were recorded.

Attributes	Like football	Dislike football	Total
Educated	55	45	100
Uneducated	125	75	200
Total	180	120	300

Test the hypothesis about independence between education and liking of football at $\alpha = 0.05$

Solution:

i. H_0: There is no association between education and liking of football.

H_1: There is association

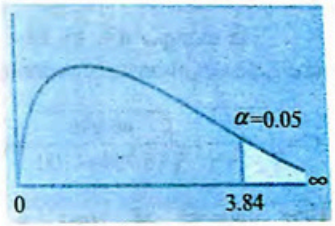
ii. $\alpha = 0.05$

iii. Test statistic to be used in this case by direct method is

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad \text{with 1 d.f}$$

iv. Calculation

$$\chi^2 = \frac{300(55\times75-45\times125)^2}{(100)(200)(180)(120)} = \frac{675}{432} = 1.5625$$



v. Critical region

Reject H_0 if $\chi_c^2 \geq \chi_{0.05,(1)}^2 = 3.84$

vi. Conclusion.

Since $\chi_c^2 = 1.5625$ falls in the acceptance χ^2 region, therefore, we accept H_0.

8.2.3 Coefficient of contingency for (r × c) contingency table

The chi-square test-statistic only decides about the independence or association of attributes in contingency tables but when H_0 is rejected it does not

tell anything about the strength of association, which we sometime need to measure. For this purpose, Karl Pearson has defined a formula for coefficient of contingency which is denoted by C and is known as Pearson's coefficient of mean-square contingency given by the relation.

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}, \quad 0\le C\le\sqrt{\frac{k-1}{k}}$$

Where χ^2 is the calculated value of chi-square test-statistic, n is the total sample size and k is the smaller one in rows and columns in number. A value of C near to "0" shows weak association and value of C near to $\sqrt{\frac{k-1}{k}}$ shows a strong association between the two attributes.

Example 8.6

Compute Pearson's coefficient of mean square contingency for the contingency table given in example 8.4.

Solution:

In example 8.4, we have $\chi^2 = 46.994$ and $n = 1100$, therefore Pearson's coefficient of mean square contingency is computed as:

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}=\sqrt{\frac{46.994}{46.994+1100}}=\sqrt{0.04097}=0.202$$

Here number of rows = number of columns = $k = 3$, so $\sqrt{\frac{k-1}{k}}=\sqrt{\frac{3-1}{3}}=\sqrt{0.6666}=0.82$. Hence the range of C is $0 \le C \le 0.82$.

Thus our computed value of C shows that the given attributes have a weak association.

8.2.4 Yate's correction for continuity

To get satisfactory results from the chi-square test in testing hypothesis about independence of attributes in contingency tables, it is necessary that expected frequency of each cell should at least be 5. If it is less, it should be added with the neighbour one to get 5 or more but in 2×2 contingency table it is not possible to combine the smaller frequency with the larger one otherwise the table will be finished. For this purpose, Frank Yate has suggested the following formula.

$$\chi^2=\sum_{i=1}^{2}\frac{\left(|O_i-e_i|-\frac{1}{2}\right)^2}{e_i}$$

This formula is known as Yate's correction for continuity and should be used only when one cell frequency is less than 5 in a 2×2 contingency table.

We also know that 2×2 contingency table is discrete frequency distribution and chi-square distribution is continuous distribution. This also needs correction for which the following formula has been suggested.

$$\chi^2=\frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

This formula should be used only when any cell expected frequency is less than 10 in a 2×2 contingency table.

8.2.5 Association versus correlation

i. Association measures the strength of relationship between two qualitative variables e.g. level of crime and education.

ii. Correlation measures the strength of relationship between two quantitative variables e.g. heights of fathers and their sons.

iii. Both are relative measures.

iv. Both measures are ranging from –1 to +1.

v. Measures of association are based on frequencies only, whereas, in correlation actual paired observations are used.

Key points

Descriptive or qualitative characteristics are called attributes.

- For the sake of convenience capital English letters A, B, C... are used to denote presence of attributes and the Greek letters $\alpha, \beta, \gamma, \ldots$ denote the absence of these attributes respectively.
- The classes A, B, AB are called positive classes because they contain the presence of attributes.
- The classes $\alpha, \beta, \alpha\beta$ are called negative classes because they contain absence of the attributes.
- For attributes class frequencies are denoted by enclosing the class symbols in parentheses.
- When a single attribute is under study then simple process of classifying the whole data in two groups is called dichotomy.
- The order of a class is determined by the number of attributes present in a class.
- The classes and class frequencies of the highest order are called the ultimate classes and ultimate class frequencies.
- Association is a statistical technique which measures the strength and direction of relationship among qualitative variables.
- $(AB) = \frac{(A)(B)}{n}$. This is known as criterion of independence for attributes.
- $Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$ is Yule's coefficient of association.
- When a population or sample is divided into many classes or categories according to an attribute is called multi-way or manifold classification.
- A table consisting of r-rows and c-columns into which the data are classified according to two attributes is called $(r \times c)$ contingency table.
- Chi-square test statistic is used to test the hypothesis about independence of attributes in the contingency tables.

Exercise

8.1 Mark the following statements as true and false.

i. Data recorded on attributes are called quantitative data.

ii. A, B, AB, ABC are called positive classes.

iii. $(B) + (\beta) = n$

iv. If $(AB) = \frac{(A)(B)}{n}$ means A and B are positively associated.

v. If $Q = 0$, then attributes are independent.

vi. Relationship between two qualitative variables is called correlation.

vii. χ^2 test has $v = (r-1)(c-1)$ df for an $(r \times c)$ contingency table.

viii. For a (4×5) contingency table the degrees of freedom for χ^2 test is 20.

ix. The range of χ^2 distribution is from 0 to ∞.

x. (2×2) contingency table contains 3-rows and 3 columns.

8.2 Fill in the suitable words in the blanks.

i. If an attribute has two classes, it is said to be _____.

ii. The number of letters representing a class determines the _____ of the class.

iii. The class represented by AB is of _____ order.

iv. The total of all frequencies n is of order _____.

v. If the attributes A and B are independent, frequency (AB) is equal to _____.

vi. If A and B are independent, Yule's coefficient of association Q is equal to _____.

vii. Association and correlation are _____.

viii. Yule's coefficient Q is ranging from _____.

ix. For an $(r \times c)$ contingency table, the sum of observed and expected frequencies must be _____.

x. For (5×6) contingency table the degree of freedom for χ^2 test is equal to _____.

8.3 Choose the correct answer.

i. Relationship between two categorical variables is called

(a) regression (b) correlation

(c) association (d) coefficient of variation

ii. The combination AB of attributes is known as the class of

(a) first order (b) second order

(c) third order (d) zero order

iii. For a 4×5 contingency table the degrees of freedom for the χ^2 is

(a) 20 (b) 16 (c) 15 (d) 12

iv. The range of Yule's coefficient of association is

(a) 0 to ∞ (b) $-\infty$ to 0

(c) 0 to 1 (d) -1 to $+1$

v. With two attributes, the total number of ultimate class frequency is

(a) two (b) four

(c) six (d) five

vi. If $(AB) < \frac{(A)(B)}{n}$, then the attributes are

(a) independent (b) positively associated

(c) negatively associated (d) no association

vii. χ^2 test statistic value varies form

(a) $-\infty$ to 0 (b) 0 to ∞ (c) $-\infty$ to ∞ (d) -1 to $+1$

viii. If A and B are independent attributes then the coefficient of association is equal to

(a) -1 (b) $+1$ (c) 0 (d) 0.5

ix. The degrees of freedom for a (2×2) contingency table will always be equal to

(a) 4 (b) 2 (c) 1 (d) 0

x. In case of consistent data, no class frequency can be

(a) positive (b) negative

(c) zero (d) one

8.4 What is meant by attributes and how they are classified?

8.5 Write short notes on the following:

i. Class symbol and class frequency

ii. Positive and negative classes

iii. Order of classes

iv. Ultimate class frequencies

8.6 Explain the following:

i. Two-way classification

ii. Association of attributes

iii. Positive and negative association

iv. Criterion of independence of attributes.

v. Yule's coefficient of association.

8.7 Distinguish between associations and correlation?

8.8 When are two attributes said to be

i. Independent

ii. Positively associated

iii. Negatively associated

8.9 Calculate the coefficient of association between extravagance in fathers and sons:

Attributes	extravagant sons	miserly sons
extravagant fathers	237	545
miserly fathers	741	235

8.10 Discuss the association when $(AB) = 256$, $(\alpha B) = 768$, $(A\beta) = 48$, $(\alpha\beta) = 144$.

8.11 Do you find any association between the tempers of brothers and sisters from the data given below?

Good natured brothers and good natured sisters = 1230

Good natured brothers and sullen sisters = 850

Sullen brothers and good nature sisters = 530

Sullen brothers and sullen sisters = 980

8.12 Investigate the association between intelligence in fathers and sons from the following data

Attributes	intelligent sons	dull sons
intelligent fathers	240	80
dull fathers	90	570

8.13 Can vaccination be regarded as a preventive measure for small-pox from the data given below?

Of 1482 persons in a locality exposed to smell-pox, 368 in all were attacked. Of 1482 persons, 343 had been vaccinated and of these 35 were attacked.

8.14 Consider the data given in the following 2 × 2 contingency table and find out whether vaccination is effective in preventing the attack of B hepatitis disease?

Attributes	attacked	not attacked
vaccinated	11	538
not vaccinated	70	464

Test by applying χ^2-test at 1% level of significance.

8.15 A random sample of size 1024 was classified according to gender and seat belt usage as shown below.

Attributes	use seat belt	don't use seat belt
male	272	192
female	276	284

Do the data suggest an association between gender and seat belt used? Use $\alpha = 0.01$

8.16 Perform chi square test of independence to decide whether smoking is a cause of lung cancer by considering the date given in the following 2×2 category table.

Attributes	cancer	no cancer
smoking	75	34
not smoking	28	112

Test at 5% level of significance.

8.17 Explain the following:

i. Manifold classification

ii. $(r \times c)$ contingency table

iii. Chi-square distribution.

8.18 Find the chi-square to test the hypothesis that there is association between height of fathers and height of sons (at $\alpha = 0.01$):

Sons \ Fathers	very tall	tall	short
very tall	600	280	300
tall	400	700	400
short	250	400	800

8.19 Can we say that education depends on sex at $\alpha = 0.05$ on the basis of a random sample of 300 persons classified in the following (2×3) contingency table:

Education / Sex	middle	SSC	college
male	30	45	75
female	75	30	45

8.20 Find the value of chi-square from the following table and test the hypothesis that there is no relation between the attributes A and B. Use $\alpha = 0.05$

Attributes	B_1	B_2	B_3
A_1	600	225	180
A_2	300	660	320
A_3	200	275	600

8.21 Discuss coefficient of mean square contingency for an $(r \times c)$ contingency table.

8.22 Find the Pearson's coefficient of mean square contingency form the following table of frequencies showing the gender of customer and mode of payment for purchasing articles on a store.

Attributes	cash	loan
male	431	5
female	291	9

8.23 Calculate the coefficient of contingency from the following data showing resemblance between hair colour and eye colour:

Hair colour / Eye colour	black	brown	red	grey
grey	1768	807	47	189
blue	946	1287	53	746
brown	115	438	6	288

Unit - 9

Design of Experiment

After studying this unit, the students will be able to

- Describe the meaning of the design of experiment.
- Explain the elements involved in designing an experiment: the experimental unit, the treatment, the replication and the response.
- Define randomization, completely randomized design.
- Give layout plan of completely randomized design.
- Identify the merits and demerits of completely randomized design.
- Know the meaning of analysis of variance.
- State the assumptions of analysis of variance.
- Describe and calculate: the total sum of squares, the treatment sum of squares, and the error sum of squares.
- Describe and calculate the degrees of freedom for: the total sum of squares, the treatment sum of squares and the error sum of squares.
- Describe and calculate the treatment mean square, the error mean square. Test the equality of means of several normal populations.

9.1 Introduction to design of experiment

In our daily life, we compare things directly by using the results, which is not good because these results are not only due to the effect of those things which we are comparing but are also influenced by some hidden factors. For example, if we want to compare the yield of two varieties of wheat, one is sown in the hot area of D.I. Khan and the second in the cold area of swat and on the basis of their production we say that one is better than the other is not justifiable because the production is surely also affected by soil, amount of water, weather, cultivation techniques, fertilizer, pesticide etc. To find the real difference between the two varieties it is reasonable to make a plan where both the varieties are given a fair chance to show their effects in the form of yield while all other factors, mentioned above are kept under control. Such a plan of experiment in which all situations except that of treatments are kept under control as much as possible, is known as design of experiments. Experimental designs are widely used for comparing treatments under identical conditions.

9.1.1 Explanation of basic terms

- **Treatment**

A thing whose effect is measured and is compared with others is called treatment. For example, to test the effect of feeds on cows, medicines on patients, dates of sowing on the yield of crop, teaching methods on students, etc. Here feeds, medicines, dates of sowing, teaching methods are treatments.

- **Experimental material**

The total material or objects on which the experiment is done is known as experimental material. For example, cows, patients, soil, students etc.

- **Experimental unit**

The smallest division of an experimental material to which a treatment is applied is called experimental unit. For example, a cow, a patient, a plot of land, a student etc.

- **Yield**

The results obtained from the experimental units after applying the treatment are called yield or responses.

- **Block**

A group of homogenous experimental units e.g. land of same fertility, students of same age, weight, I.Q, etc. to which all treatments are assigned at random is called a block. For soil fertility blocks are always made perpendicular to variation. Generally, blocks are made by uniformity trial. Each treatment is given a chance in each block and each appears in a block only once.

- **Uniformity trial**

It is a trial or an experiment in which same treatment is given to all experimental units and then experimental units of the same performance is consider as a block. For example, same test is given to whole class then all the students who get first division is a block, all the students getting second division is a block, all third division holders is a block. Uniformity trial is used only for making blocks.

- **Extraneous factor**

The responses from all experimental units receiving the same treatment may not be identical even if the experiment is performed under similar conditions. These responses are influenced not only by treatments but also by other factors as well, some of which can be controlled while there are some over which there is no control or very little control. For example, natural differences between the experimental units like heterogeneity of soil, I.Q of students, climatic factors etc. All such factors which are not in the control of researcher are called uncontrolled or extraneous factors.

- **Experimental error**

The error caused by extraneous factors which are beyond the control of human approach is known as experimental error. It is a major problem for experimenter and is a mask on the true effect of the treatments because the observed difference

in treatments is a sum of the true difference of the treatments plus due to the experimental error. Design of experiment is actually a strategy for controlling the experimental error in order to bring out the real difference among the treatments.

9.1.2 Basic requirements or principles of a good experimental design

An experimental design can reduce the experimental error only if it follows the three requirements of a good design namely, randomization, replication and local control.

(i) Randomization

The allocation of treatments to experimental units in such a manner that an experimental unit has equal chance of receiving any of the treatments is called randomization. It is usually done by (a) using lottery method (b) using random number table method (c) using any computer package which may perform randomization.

- Advantages of randomization are:
  a) Eliminates the human biases.
  b) Introduces the independence in the assignment of treatments to the experimental units which in turn creates independence amongst the observations, required for the validity of F-test.

(ii) Replication

Repetition of a treatment on a number of experimental units in an experiment is known as replication of the treatment. Replications are essential because in all experiments, there are great variations in fertility of the experimental units and all the treatments do not get equal chance of experiencing every type of environment in the field. If a treatment is allotted only once and it goes to more fertile experimental unit, will be in more advantageous position than those which are applied to less fertile experimental units. This type of variation can be eliminated by using replication process which improve the precision of an experiment and provides a valid estimate of the experimental error.

(iii) Local control

Local control is a procedure which maintains greater homogeneity of experimental units within a block of an experiment. For example, soil fertility of field is a factor which affects the plant growth and yield. All the plots having the same soil fertility should constitute a block. The soil fertility of land can be assessed by conducting a uniformity trial on the field prior to actual field experiment. The treatments should be assigned to the blocks in equal number of times. Local control also called error control reduces the experimental error. Note that all these three principles contribute a lot in increasing the efficiency of design.

9.2 Completely randomized design

Experimental design, in which the treatments are allocated completely at random to the experimental units, is called completely randomized design (CRD). In this design all the experimental units which are homogeneous, as much as possible, are taken as a single group. Any number of treatments and any number of units per treatment may be used. It is the simplest design.

9.2.1 Experimental layout for CRD

The layout of an experiment is the actual placement of the treatments on the experimental units. To explain the procedure of randomization let us consider four treatments, each replicated three times.

Step 1: Determine the total number of plots $n = t \times r = 4 \times 3 = 12$

Step 2: Assign a plot number to the experimental plots or units as follows:

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

Step 3: Assign treatments to the experimental units. The randomization both by lottery and random number table methods is explained as follows:

(i) Lottery method

Make 12 slips of paper. Write T_1 on 3 slips, T_2 on 3 slips, T_3 on 3 slips, T_4 on 3 slips and place them in a bowl and mix them thoroughly. Draw the slips one at a time without replacement and allot to the above experimental units serially from 1 to 12

(ii) Random number table method

Take any random number table, start from anywhere and take twelve two-digit random numbers. Exclude the number which is greater than 12 or which is repeated. Let the random numbers selected are: 03 02 04 06 01 07 11 08 10 12 09 05, so allot treatment T_1 to (03, 02, 04) experimental units, T_2 to (06, 01, 07) experimental units, T_3 to (11, 08, 10) and T_4 to (12, 09, 05). The layout will appear as:

1 T_2	2 T_1	3 T_1	4 T_1
5 T_4	6 T_2	7 T_2	8 T_3
9 T_4	10 T_3	11 T_3	12 T_4

• Merits of completely randomized design

i. Its layout is very easy.
ii. There is complete flexibility in this design i.e. any number of treatments and replicates for each treatment can be taken.
iii. Whole experimental material can be utilized in this design.
iv. This design yields maximum degrees of freedom for experimental error.
v. The analysis of data, both for equal and unequal number of replications is simplest as compared to any other design.
vi. Missing observation(s) creates no problem in analysis of data, nor efficiency of the design affected.
vii. CRD is suitable in situations where a fraction of the units is likely to be destroyed during experimentation or is likely to fail to respond.

• Demerits of completely randomized design

i) It is difficult to find homogeneous experimental units in all respects.
ii) The design is suitable for a small number of treatments.
iii) CRD is seldom suitable for field experiments as compared to other experimental designs.

9.3 Analysis of variance

We have learnt that Z-test and t-test are used for testing hypothesis about population mean or equality of two population means. Analysis of variance (ANOVA) is also a hypothesis testing procedure that is used to compare three or more populations or treatments means. ANOVA procedure provides greater flexibility in designing experiments, analysis and interpretation of experimental results to the researchers. This procedure was developed by the British statistician Sir Ronald A. Fisher.

Definition

Measurements are taken on each experimental unit according to characteristic of interest and its variance is calculated. This total variance is due to various factors involved in the experiment. "ANOVA is a technique which split the total variance in to meaningful component variances; each gives an estimate of the population variance. The ratio of two component variances is distributed as F with corresponding degrees of freedom." Thus ANOVA enables one to know whether the variance due to a component factor is significantly more than the variance due to experimental error or not and hence decision is made about the null hypothesis.

9.3.1 Assumptions for analysis of variance

Analysis of variance technique is based on the following assumptions.

i) Independence
The samples are drawn randomly and are independent of other samples.

ii) Normality
The populations under study have distributions which are assumed to be normal having means $\mu_1, \mu_2, \ldots \mu_k$.

iii) Homogeneity

All populations under study have equal variance i.e. $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 = \sigma^2$

iv) Additivity

The random observations X_{ij} can be expressed as the sum of means μ_j and the error terms ε_{ij} as $X_{ij} = \mu_j + \varepsilon_{ij}$

Note that in practice minor deviations from the assumptions is tolerated but serious violation of these assumptions will damage the ANOVA procedure.

9.3.2 One way analysis of variance (equal sample size case)

The observed data of k samples collected for k treatments, each of size r such that $rk = n$ can be arranged in tabular form as:

Treatments / Observations	1	2	...	j	...	k	Total
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1k}	
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2k}	
⋮	⋮						
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ik}	
⋮	⋮						
r	x_{r1}	x_{r2}	...	x_{rj}	...	x_{rk}	
Total T_j	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$T_{.k}$	$T_{..}$ (Grand total)
Means $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$	...	$\bar{x}_{.j}$	...	$\bar{x}_{.k}$	$\bar{x}_{..}$ (Grand mean)
Sum of squares	$\sum_{i=1}^{r}(x_{i1}-\bar{x}_{.1})^2$	$\sum_{i=1}^{r}(x_{i2}-\bar{x}_{.2})^2$		$\sum_{i=1}^{r}(x_{ij}-\bar{x}_{.j})^2$		$\sum_{i=1}^{r}(x_{ik}-\bar{x}_{.k})^2$	$\sum_{j=1}^{k}\sum_{i=1}^{r}(x_{ij}-\bar{x}_{..})^2$ (Total sum of square)

9.3.3 Components of the total sum of squares (equal sample size)

In one way analysis of variance the total variation present in the observed data may be due to two meaningful components i.e. (i) variation among treatments (samples) (ii) variation with in the samples, so the T.S.S is divided in to two components as:

$$\text{T.S.S} = \sum_{j=1}^{k}\sum_{i=1}^{r}(x_{ij} - \bar{x}_{.})^2$$

$$= \text{within SS} + \text{Treatment SS}$$

To obtain unbiased estimates for σ^2 divide both sides by respective degrees of freedom to have

$$S_T^2 = S_e^2 + S_{Tr}^2$$

Or MST = MSE + MSTr

The F-Test is defined as $F = \dfrac{\text{MSTr}}{\text{MSE}}$ with ($v_1 = k-1$ and $v_2 = n-k$) df. which is used for testing H_0: $\mu_1 = \mu_2 = ... = \mu_k$ verses

H_1: Not all means are equal.

H_0 is rejected when, $F_c \geq F_{\alpha,(v_1,v_2)}$, whereas $F_{\alpha,(v_1,v_2)}$ is obtained from the percentage points of the F-distribution, Table 9.1.

9.3.4 Components of the total degrees of freedom

When the total sum of squares is partitioned into component parts then the total degrees of freedom can also be partitioned into component parts accordingly as below:

$$(n-1) = (n-k) + (k-1)$$

Total df = within df + between df

9.3.5 ANOVA table

For convenience all sources of variation (S.O.V), degrees of freedom (df), sum of squares (S.S), mean squares (M.S) are presented in tabular form, called ANOVA table given below:

ANOVA table

S.O.V	df	S.S	M.S	F_{cal}
Treatment	$k-1$	T_rSS	$\frac{TrSS}{k-1}=MSTr$	$F_c=\frac{MSTr}{MSE}$
Error	$n-k$	ESS	$\frac{ESS}{n-k}=MSE$	—
Total	$n-1$	TSS	$\frac{TSS}{n-1}=MST$	—

This table is used for testing hypothesis H_0: $\mu_1=\mu_2=\cdots=\mu_k$ in one way analysis of variance.

- **Formulas for sum of squares**

Correction factor: $C.F=\frac{T_{..}^2}{n}$

Total sum of squares: $TSS=\sum_{j=1}^{k}\sum_{i=1}^{r}x_{ij}^2-C.F$

Treatment sum of squares: $TrSS=\frac{\sum_{j=1}^{k}T_{.j}^2}{r}-C.F$

Error sum of squares: $ESS=TSS-T_rSS$

9.3.6 General procedure for testing hypothesis in case of one way ANOVA (equal sample size case)

i. H_0: $\mu_1=\mu_2=\cdots=\mu_k$

H_1: Not all means are equal

ii. Choose an appropriate level of significance

iii. Test statistic to be used in this case is

$F_c=\frac{MSTr}{MSE}$ with $(v_1=k-1, v_2=n-k)$ d.f

iv. Computations.

v. Critical region: reject H_0 if $F_{cal}\geq F_{tab}$

vi. Conclusion

Table 9.1 (Percentage points of the F-distribution)

At $\alpha = \begin{cases} 5\% \\ 1\% \\ 10\% \end{cases}$ Level of Significance

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	11
1	161.4 4052 39.1	199.5 4999 49.5	215.7 5403 53.6	224.6 5625 55.8	230.2 5859 58.2	234.0 5859 58.2	236.8 5928 58.9	238.9 5981 59.4	240.5 6022 59.9	241.9 6056 60.2	243 6081 60.4
2	18.51 98.50 8.53	19.00 99.00 9.0	19.16 99.17 9.16	19.25 99.25 9.24	19.30 99.30 9.29	19.33 99.33 9.33	19.35 99.36 9.35	19.35 99.36 9.35	19.38 99.39 9.38	19.40 99.40 9.39	19.4 99.4 9.40
3	10.13 32.12 5.54	9.522 30.82 5.46	9.27 29.46 5.39	9.12 28.71 5.34	9.013 28.24 5.31	8.94 27.91 5.28	8.89 27.67 5.27	8.89 27.67 5.27	8.81 27.35 5.24	8.79 27.23 5.23	8.75 27.1 5.22
4	7.709 21.20 4.45	6.94 18.00 4.32	6.59 16.69 4.19	6.39 15.98 4.11	6.26 15.52 4.05	6.16 15.21 4.01	6.09 14.98 3.98	6.09 14.98 3.98	5.99 14.66 3.94	5.96 14.55 3.92	5.94 14.4 3.91
5	6.608 16.26 4.06	5.79 13.27 3.78	5.41 12.06 3.62	5.19 11.39 3.52	5.05 10.97 3.45	4.95 10.67 3.40	4.88 10.46 3.37	4.82 10.29 3.34	4.77 10.16 3.32	4.47 10.05 3.30	4.70 9.96 3.29
6	5.987 13.75 3.78	5.14 10.92 3.46	4.76 9.78 3.29	4.53 9.15 3.18	4.39 8.75 3.11	4.29 8.47 3.05	4.21 8.26 3.01	4.15 8.10 2.98	4.09 7.98 2.96	4.06 7.87 2.94	4.03 7.79 2.92
7	5.59 12.25 3.59	4.74 9.55 3.26	4.35 8.45 3.07	4.12 7.85 2.96	3.97 7.46 2.88	3.87 7.19 2.83	3.79 6.969 2.78	3.73 6.84 2.75	3.68 6.72 2.72	3.64 6.62 2.70	3.60 6.54 2.69
8	5.318 11.26 3.46	4.46 8.65 3.11	4.07 3.59 2.92	3.84 7.01 2.81	3.69 6.63 2.73	3.58 6.37 2.67	3.50 6.18 2.62	3.44 6.03 2.59	3.39 5.91 2.56	3.35 5.81 2.54	3.31 5.73 2.52
9	5.117 10.56 3.26	4.26 8.02 3.01	3.86 6.99 2.81	3.36 6.42 2.69	3.48 6.06 2.61	3.37 5.80 2.55	3.29 5.61 2.51	3.23 5.47 2.47	3.18 5.35 2.44	3.14 5.26 2.42	3.10 5.18 2.40
10	4.965 10.04 3.28	4.10 7.56 2.92	3.71 6.55 2.73	3.48 5.99 2.61	3.33 5.64 2.52	3.22 5.39 2.46	3.14 5.20 2.41	3.07 5.06 2.38	3.02 4.94 2.35	2.98 4.85 2.32	2.94 4.77 2.30
11	4.84 9.65 3.23	3.98 7.21 2.86	3.59 6.22 2.66	3.36 5.67 2.54	3.20 5.32 2.45	3.09 5.07 2.39	3.01 4.89 2.34	2.95 4.74 2.30	2.89 4.63 2.27	2.85 4.54 2.25	2.82 4.46 2.23
12	4.75 9.33 3.18	3.89 6.93 2.81	3.49 5.95 2.61	3.26 5.41 2.48	3.11 5.06 2.39	2.99 4.82 2.33	2.91 4.64 2.28	2.85 4.49 2.24	2.79 4.39 2.21	2.75 4.29 2.19	2.72 4.22 2.17
13	4.67 9.07 3.14	3.81 6.70 2.76	3.41 5.74 2.56	3.18 5.21 2.43	3.03 4.86 2.35	2.91 4.62 2.28	2.83 4.44 2.23	2.77 4.30 2.20	2.71 4.19 2.16	2.67 4.10 2.14	2.63 4.02 2.12
14	4.60 8.86 3.80	3.74 6.52 2.73	3.34 5.56 2.52	3.11 5.04 2.39	2.96 4.69 2.31	2.85 4.46 2.24	2.76 4.28 2.19	2.69 4.14 2.15	2.65 4.03 2.12	2.60 3.94 2.10	2.57 3.88 2.08
15	4.54 8.68 3.07	3.68 6.36 3.07	3.29 5.42 2.70	3.06 4.89 2.49	2.90 4.96 2.36	2.79 4.32 2.27	2.71 4.14 2.21	2.64 4.00 2.16	2.58 3.89 2.12	2.54 3.81 2.06	2.51 3.73 2.04
16	4.49 8.53 3.05	3.63 6.23 2.67	3.24 5.29 2.46	3.01 4.77 2.33	2.85 4.44 2.24	2.74 4.20 2.18	2.66 4.03 2.13	2.59 3.89 2.09	2.54 3.78 2.06	2.49 3.69 2.03	2.46 3.62 2.01
17	4.45 8.40 3.03	3.59 6.11 2.64	3.20 5.18 2.44	2.96 4.67 2.31	2.81 4.34 2.22	2.70 1.10 2.15	2.61 3.93 2.10	2.55 3.79 2.06	2.49 3.68 2.03	2.45 3.59 2.00	2.41 3.52 1.98
18	4.41 8.29 3.01	3.55 6.01 2.62	3.16 5.09 2.42	2.93 4.58 2.29	2.77 4.25 2.20	2.66 4.01 2.13	2.58 3.84 2.08	2.51 3.71 2.04	2.46 3.60 2.00	2.41 3.51 1.98	2.37 3.43 1.96
19	4.38 8.18 2.99	3.52 5.93 2.61	3.13 5.01 2.40	2.90 4.50 2.27	2.74 4.17 2.18	2.63 3.94 2.11	2.54 3.77 2.06	2.48 3.63 2.02	2.42 3.52 1.98	2.38 3.43 1.96	2.34 3.36 1.94
20	4.35 8.10 2.97	3.49 5.85 2.59	3.10 4.94 2.38	2.87 4.43 2.25	2.71 4.10 2.16	2.60 3.87 2.09	2.51 3.70 2.09	2.45 3.56 2.00	2.39 3.46 1.96	2.35 3.37 1.94	2.31 3.29 1.92
25	4.24 7.77 2.92	3.39 5.57 2.53	2.99 4.68 2.32	2.76 4.18 2.18	2.60 3.86 2.09	2.49 3.63 2.02	2.40 3.46 1.97	2.34 3.32 1.93	2.28 3.22 1.89	2.24 3.13 1.87	2.20 3.06 1.85

30	4.17 7.56 2.88	3.32 5.39 2.49	2.92 4.51 2.28	2.69 4.02 2.14	2.53 3.70 2.05	2.42 3.47 1.98	2.33 3.30 1.93	2.27 3.17 1.88	2.21 3.07 1.85	2.16 2.98 1.82	2.13 2.91 1.80
35	4.12 7.42 2.86	3.27 5.27 2.47	2.88 4.40 2.26	2.64 3.91 2.12	2.49 3.59 2.03	2.37 3.37 1.96	2.29 3.20 1.90	2.22 3.07 1.86	2.16 2.97 1.82	2.12 2.87 1.79	2.08 2.81 1.77
40	4.08 7.31 2.84	3.23 5.18 2.44	2.84 4.31 2.23	2.61 3.83 2.09	2.45 3.54 2.00	2.34 3.29 1.93	2.25 3.12 1.87	2.18 2.99 1.83	2.12 2.89 1.79	2.08 2.80 1.76	2.04 2.73 1.74
45	4.06 7.23 2.82	3.21 5.11 2.43	2.82 4.25 2.22	2.58 3.77 2.08	2.43 3.46 1.39	2.31 3.23 1.92	2.23 3.07 1.86	2.16 2.94 1.82	2.10 2.83 1.78	2.05 2.74 1.75	2.01 2.73 1.73
50	4.03 7.17 2.81	3.18 5.06 2.41	2.79 4.20 2.20	2.56 3.72 2.06	2.40 3.41 1.97	2.29 3.19 1.90	2.20 3.02 1.84	2.13 2.89 1.80	2.07 2.79 1.76	2.03 2.70 1.73	1.99 2.63 1.71
60	4.00 7.08 2.79	3.15 4.98 2.39	2.76 4.13 2.18	2.53 3.65 2.04	2.37 3.34 1.95	2.25 3.12 1.87	2.17 2.95 1.82	2.10 2.82 1.77	2.04 2.72 1.74	1.99 2.63 1.71	1.95 2.56 1.69
70	3.98 7.01 2.78	3.13 4.92 2.38	2.74 4.08 2.17	2.50 3.60 2.03	2.35 3.29 1.94	2.23 3.07 1.86	2.14 2.91 1.81	2.07 2.78 1.76	2.02 2.67 1.73	1.97 2.59 1.69	1.93 2.51 1.67
80	3.96 6.96 2.77	3.11 4.88 2.37	2.72 4.04 2.15	2.19 3.56 2.02	2.33 3.26 1.92	2.21 3.01 1.85	2.13 2.87 1.79	2.06 2.74 1.75	2.00 2.64 1.71	1.95 2.55 1.68	1.91 2.48 1.65
90	3.95 6.93 2.77	3.10 4.85 2.37	2.71 4.01 2.15	2.47 3.54 2.01	2.32 3.23 1.92	2.20 3.01 1.84	2.11 2.84 1.79	2.04 2.72 1.74	1.99 2.61 1.71	1.94 2.52 1.67	1.90 2.45 1.65
100	3.94 6.60 2.76	3.09 4.82 2.36	2.70 3.98 2.14	2.46 3.51 2.00	2.31 3.21 1.91	2.19 2.99 1.83	2.10 9.82 1.78	2.03 2.69 1.73	1.97 2.59 1.70	1.93 2.50 1.66	1.89 2.43 1.64

Example 9.1

The attention spans in minutes were randomly observed by a teacher from a group of 15 first year students during a morning reading period and are shown in the following table.

No breakfast	Light breakfast	Full breakfast
8	14	10
7	16	12
9	12	16
13	17	15
10	11	12

Do the data indicate that the average attention spans of the three treatments are equal? Use $\alpha = 0.05$

Solution:

i. H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Not all means are equal

ii. $\alpha = 0.05$

iii. Test statistic to be used is

$$F_c = \frac{MST_r}{MSE} \sim F\text{-distribution with } v_1 = k\text{-}1, v_2 = n\text{-}k \text{ d.f}$$

iv. Calculation:

	x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	Total
	8	64	14	196	10	100	
	7	49	16	256	12	144	
	9	81	12	144	16	256	
	13	169	17	289	15	225	
	10	100	11	121	12	144	
$T_{.j}$	47		70		65		$182 = T_{..}$
$T^2_{.j}$	2209		4900		4225		$11334 = \sum_{j=1}^{3} T_{.j}^2$
$\sum_{i=1}^{5} x_{.j}^2$		463		1006		869	$\rightarrow 2338 = \sum_{i=1}^{5}\sum_{j=1}^{3} x_{ij}^2$

$$C.F = \frac{T^2_{..}}{n} = \frac{(182)^2}{15} = 2208.2667$$

$$TSS = \sum_{j=1}^{3}\sum_{i=1}^{5} x_{ij}^2 - C.F = 2338 - 2208.2667 = 129.7333$$

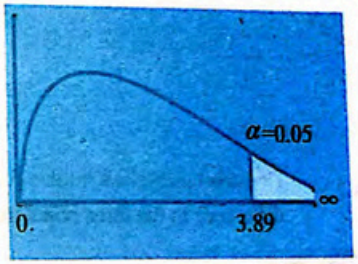
$$T_rSS = \frac{\sum_{j=1}^{3} T_{.j}^2}{r} - C.F = \frac{11334}{5} - 2208.2667 = 58.5333$$

$$ESS = TSS - T_rSS = 129.7333 - 58.5333 = 71.2$$

ANOVA table

S.O.V	d.f	S.S	M.S	F
Treatments	$3-1=2$	58.5333	$\frac{58.5333}{2}=29.27$ MSTr	$\frac{MSTr}{MSE}=\frac{29.27}{5.93}=4.93$
Error	$15-3=12$	71.2	$\frac{71.2}{12}=5.93$ MSE	—
Total	$15-1=14$	129.7333	—	—

v. Critical region



Reject H_0 if $F_c \geq F_{tab}$ whereas from F-table 9.1

$$F_{tab} = F_{\alpha,(\nu_1,\nu_2)} = F_{0.05,(2,12)} = 3.89$$

vi. Conclusion:

Since $F_c = 4.93$ falls in the rejection region, therefore, we reject H_0 and conclude that average attention span of students is different for the three treatments (types of breakfast).

9.3.7 One way ANOVA (unequal sample sizes)

Practically sample sizes may not always be equal. One way ANOVA for equal sample sizes will still be used with minor changes in the sum of square formulas as follows:

$$n = r_1 + r_2 + \ldots\ldots + r_k = \sum_{j=1}^{k} r_j$$

$$TSS = \sum_{j=1}^{k}\sum_{i=1}^{rj} x^2_{ij} - C.F$$

$$T_rSS = \sum_{j=1}^{k} \frac{T^2_{.j}}{r_j} - C.F$$

Example 9.2

The following observations are collected using a completely randomized design:

Sample1	Sample 2	Sample3
3	4	2
2	3	0
4	5	2
3	2	1
2	5	

Construct an ANOVA table for the data and determine whether there is a difference in the three population means. Use $\alpha = 0.01$

Solution:

i. H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Not all means are equal

ii. $\alpha = 0.01$

iii. Test statistic to be used is

$$F = \frac{MSTr}{MSE} \text{ with } (v_1 = k-1, v_2 = n-k) \text{ d.f}$$

iv. Calculation:

	x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	Total
	3	9	4	16	2	4	
	2	4	3	9	0	0	
	4	16	5	25	2	4	
	3	9	2	4	1	1	
	2	4	5	25			
$T_{.j}$	14		19		5		$38 = T_{..}$
$T^2_{.j}$	196		361		25		
$\sum_{i=1}^{5} x_j^2$		42		79		9	$130 = \sum_j \sum_i x_{ij}^2$

$n = r_1 + r_2 + r_3 = 5 + 5 + 4 = 14$

$$C.F = \frac{T_{..}^{\ 2}}{n} = \frac{(38)^2}{14} = 103.1429$$

$$TSS = \sum_{j=1}^{3}\sum_{i-1}^{r_j} x_{ij}^2 - C.F = 130 - 103.1429 = 26.8571$$

$$T_rSS = \sum_{j=1}^{3} \frac{T^2_{.j}}{r_j} - C.F$$

$$= \left[\frac{T^2_{.1}}{r_1} + \frac{T^2_{.2}}{r_2} + \frac{T^2_{3}}{r_3}\right] - C.F$$

$$= \left[\frac{196}{5} + \frac{361}{5} + \frac{25}{4}\right] - C.F$$

$$= [39.2 + 72.2 + 6.25] - 103.1429 = 14.5071$$

$$ESS = TSS - T_rSS$$

$$= 26.8571 - 14.5071 = 12.35$$

Put all these values in ANOVA table we have

ANOVA table

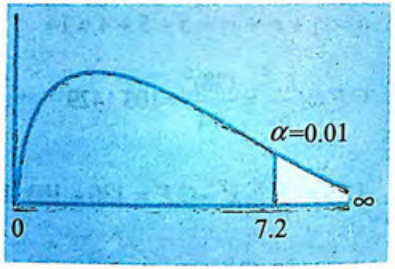
S.O.V	df	S.S	M.S	F
Treatments	2	14.5071	7.2536	6.461
Error	11	12.35	1.1227	—
Total	13	26.8571	—	—

v. Critical region

Reject H_0 if $F_c \geq F_{tab}$ whereas,

$F_{tab} = F_{\alpha,(v_1,v_2)} = F_{0.01,(2,11)} = 7.2$

(From F- table 9.1)



vi. Conclusion:

Since our computed value of $F_c = 6.461$ lies in the acceptance region, therefore, we accept H_0.

Key points

- A plan of experiment in which all situations except that of treatments are kept under control as much as possible, is known as design of experiment.
- Anything whose effect is measured and is compared with others is called treatment.
- The total material or objects on which the experiment is done is known as experimental material.
- The smallest division of an experimental material to which a treatment is applied is called experimental unit.
- The results obtained from the experimental units are called yield or responses.
- A group of homogenous experimental units e.g. land of same fertility, students of same age, weight, I.Q, etc. to which all treatment are assigned at random is called a block.
- Uniformity trial is used only for making blocks.
- All factors which are not in the control of a researcher are called uncontrolled or extraneous factors.
- The error which arises due to the extraneous factors is called experimental error.
- Design of experiment is actually a strategy for controlling the experimental error in order to bring out the real difference among the treatments.
- The allocation of treatments to experimental units in such a manner that an experimental unit has equal chance of receiving any of the treatments is called randomization.
- Repetition of a treatment on a number of experimental units in an experiment is known as replication of the treatment.
- Local control also called error control reduces the experimental error.
- Experimental design in which the treatments are allocated randomly to the experimental units is called completely randomized design
- ANOVA is a technique which split the total variance into meaningful component variances; each gives an estimate of the population variance. The ratio of two component variances is distributed as F with corresponding degrees of freedom.

Exercise

9.1 Read the following statements carefully and indicate which statement is true or false.

i. F-distribution is a ratio of sample variances.
ii. ANOVA is a useful technique for testing hypothesis about several means.
iii. In ANOVA sample sizes must always be equal.
iv. In ANOVA it is assumed that the populations from which samples are drawn are normally distributed
v. F-test is used in ANOVA to test the null hypothesis about several means.
vi. The range of F-distribution is from 0 to ∞.
vii. The plan of an experiment which controls all factors, as much as possible except the treatment is called design of experiment.
viii. Larger the experimental error, more efficient is the design.
ix. A completely randomized design is used when all experimental units are homogeneous.
x. Missing observation in a CRD creates a serious problem.

9.2 Fill in the blanks.

i. A subject receiving a treatment in an experiment is called _____.
ii is The allocation of treatments to experimental units with equal probability known as _____.
iii. The number of times a treatment is repeated in an experiment is called its _____.
iv. Experimental error is the error caused by_____.
v. Smaller the experimental error, more _____is the design.
vi. ANOVA splits the total variance is into _____.
vii. Missing observation in a completely randomized design creates _____.
viii. Error sum of squares _____ be negative.
ix. The average performance of a treatment is better reflected through _____.
x. The designs of experiments were originated mainly for _____.

9.3 Choose the correct answer.

i. An experimental design is
(a) An architect (b) a map
(c) a plan of experiment (d) all of the above

ii. Randomization is a process in which the treatments are allocated to the experimental units:
(a) In a sequence (b) at the will of the investigator
(c) With equal probability (d) with unequal probability

iii. Replication in an experiment means:
(a) The number of times a treatment occurs in an experiment
(b) The numbers of blocks
(c) Total number of treatments
(d) The reduction of blocks

iv. Local control is a method to maintain
(a) Homogeneity among blocks (b) Homogeneity within blocks
(c) both (a) and (b) (d) all of the above

v. Experimental error is due to
(a) Variation in treatment effects (b) extraneous factors
(c) Experimenter's mistake (d) lack of experience

vi. Completely randomized designs are mostly used in
(a) Pot experiments (b) Experiments on animals
(c) Field experiments (d) all of the above

vii. An experimental unit in a research work is
(a) A patient (b) A field plots
(c) An animal (d) All of the above

viii. Factors fertilizer, date of sowing and breeds are called

(a) Replicates (b) experimental unit

(c) Treatments (d) All of the above

ix. Local control in the field is maintained through

(a) Natural factors (b) Randomization

(c) Replication (d) Uniformity trials

x. Randomization in an experiment helps the researcher to eliminate

(a) Dependence among observations

(b) Systematic influences

(c) Human biases

(d) all of the above

9.4 Describe the design of an experiment in your own words.

9.5 Define the following.

i. Experimental material

ii. Experimental unit

iii. Treatment

iv. Uniformity trial

v. Block

9.6 What do you understand by randomization, replication and local control in experimental design?

9.7 i. What is meant by experimental design?

ii. What are the basic principles of design of experiment?

9.8 Discuss and define (a) Extraneous factor (b) experimental error.

9.9 Discuss the need and utility of planning a statistical experiment.

9.10 What is a completely randomized design? What are the merits and demerits of a completely randomized design?

9.11 Explain the experimental layout for a completely randomized design using 3 treatment and 15 experiment plots.

9.12 Write a short note on analysis of variance.

9.13 What is meant by analysis of variance? What are the assumptions under which this technique is applied?

9.14 What do you understand by?

i. Variance among samples

ii. Variance within samples

iii. Total variation

9.15 Given the data below, perform the analysis of variance and test the hypothesis that the means of the three populations are equal. Let $\alpha = 0.05$

X_1	X_2	X_3
13	18	17
14	19	20
16	12	8
17	15	11

9.16 Four salesmen were posted in different areas by a company. The numbers of units of commodity "x" sold by them are as follow:

A	28	23	20	29
B	30	32	25	21
C	35	28	23	18
D	19	21	15	25

Is there a significant difference in the performance of these salesmen?

9.17 The following data gives the figures of production of rice of three varieties A, B, C of rice shown in 12 plots

A	18	15	24	23
B	19	23	24	18
C	16	19	31	22

Carry out the analysis of variance and test 5% level of significance that is there a significant difference among varieties?

9.18 The three samples below have been obtained from normal populations with equal variances. Test the hypothesis at 5% level of significance that the population means are equal.

X_1	X_2	X_3
10	5	9
8	7	12
7	10	13
14	9	12
11	9	14

9.19 A test was given to 5 students chosen at random from the first year statistics class each of the three colleges in Peshawar. Their scores were found as follow.

A	50	80	90	70	60
B	40	50	70	40	50
C	70	60	60	50	60

Perform analysis of variance and show if there is any significant difference among the score of students in the three colleges. Use $\alpha = 0.05$

9.20 Test the hypothesis that no differences exist among the four treatments at $\alpha = 0.05$.

Sample 1	Sample 2	Sample 3	Sample 4
5	3	4	8
4	6	11	18
4	4	8	14
11	6	6	27

9.21 Given the following data obtained from a completely randomized design with four treatments;

X_1	X_2	X_3	X_4
12.4	14.4	10.2	6.1
20.9	9.0	13.2	5.8
10.1	23.7	5.1	4.8
4.2			1.5

Analyse the given data and draw conclusion about the equality of treatment effects. Use $\alpha = 0.05$

9.22 The following data were obtained by using completely randomized design

Sample 1	6	8	10	8
Sample 2	8	10	9	
Sample 3	10	8		
Sample 4	9	10	7	8
Sample 5	8	10	12	

Construct an ANOVA table for the data. Test $\alpha = 0.05$ that five population means are equal.

9.23 Compare the following random samples.

Sample 1: 17 19 4 9 10 11

Sample 2: 12 15 6 8 10 11 12

Sample 3: 20 23 9 13 15

Perform ANOVA and test the hypothesis at $\alpha = 0.05$ that the samples came from populations having same means.

9.24 The results shown in the following table were obtained through completely randomized design.

A	2	1	0	2	
B	2	5	3	4	5
C	3	2	3	2	4

Test at 1% level of significance that there is no difference in the means of the three populations.

ANSWERS

Exercise-1

1.1 (i) T (ii) F (iii) F (iv) T (v) T
(vi) F (vii) T (viii) T (ix) T (x) F

1.2 (i) impossible (ii) sample space (iii) permutations
(iv) 0 to 1 (v) gambling (vi) addition
(vii) $P(A) + P(B)$ (viii) independent (ix) $1 - P(A)$ (x) $\frac{6}{36}$

1.3 (i) c (ii) d (iii) c (iv) d (v) b
(vi) c (vii) d (viii) b (ix) c (x) b

1.5 (i) 40320 (ii) 6375600 (iii) 10626 (iv) 3876
(v) 167960

1.7 40320, 720

1.8 60

1.9 (i) 4989600 (ii) 15135120 (iii) 19958400
(iv) 4989600 (v) 50400

1.10 330

1.15 0.50

1.16 $\frac{5}{36}$

1.17 (i) $\frac{1}{8}$ (ii) $\frac{3}{8}$ (iii) $\frac{7}{8}$

1.18 0.0045

1.19 (i) $\frac{13}{52}$ (ii) $\frac{26}{52}$ (iii) $\frac{4}{52}$

1.20 (i) $\frac{1}{3}$ (ii) 0.133

1.21 (b) 0.92

1.22 $\frac{8}{36}$

1.24 $\frac{1}{4}$

-1.25 (i) 0.0059 (ii) 0.0045

1.26 (b) $\frac{1}{3}$

1.27 (i) 0.12 (ii) 0.6 (iii) 0.68

1.28 $\frac{3}{4}$

1.29 (i) $\frac{1}{210}$ (ii) $\frac{209}{210}$

1.30 (a) $p=\frac{1}{2}, d=1$ (b) $p=\frac{1}{4}, d=\frac{1}{3}$

Exercise-2

2.1 (i) T (ii) F (iii) T (iv) F (v) T
(vi) F (vii) T (viii) T (ix) F (x) T

2.2 (i) finite (ii) unity (iii) $\frac{5}{21}$ (iv) random variable
(v) discrete & continuous (vi) joint distribution (vii) one
(viii) $E(X)E(Y)$ (ix) S.D(X)+S.D(Y) (x) $\frac{2}{3}$

2.3 (i) b (ii) b (iii) b (iv) c (v) d
(vi) c (vii) c (viii) b (ix) a (x) d

2.6 (i) continuous (ii) continuous (iii) discrete
(iv) continuous (v) continuous (vi) discrete
(vii) continuous (viii) continuous (ix) discrete
(x) discrete

2.8

X	0	1	2
$p(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

2.9

X	0	1	2	3
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

2.10 (i) 0.3 (ii) 0.6 (iii) 0.8

2.11

X	−3	−1	1	3
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

2.12

X	−5	−4	−3	−2	−1	0	1	2	3	4	5
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

2.13 No, because sum of probabilities is not equal to one

2.14 (i) $\frac{3}{8}$ (ii) $\frac{1}{2}$ (iii) $\frac{3}{4}$

2.15 (b) (i) 3 (ii) 0.125 (iii) 0.019

2.17 (i) 0.5 (ii) 3.5 (iii) 6

2.18 0.55, 1.35, 1.16

2.19 (i) 2.6 (ii) 1.34 (iii) 1.16

2.20 (i) 20 (ii) $\frac{2}{3}$ (iii) $\frac{2}{63}$ (iv) 0.178

2.21 Mean $=\frac{4}{9}$ Variance $=\frac{13}{162}$, S.D = 0.2833

2.22 $E(X)=2$, $E(Y)=0.5$, $E(X+Y)=2.5$, $E(XY)=1$

2.23 (i) Yes, X & Y are independent (ii) 35, 36.8, 71.8, 1288

2.24 (i) 0.41 (ii) 0.24 (iii) 0.65 (iv) 0.65

Exercise-3

3.1 (i) F (ii) T (iii) F (iv) T (v) T
(vi) F (vii) T (viii) T (ix) T (x) T

3.2 (i) two (ii) one (iii) n (iv) n and p
(v) np, npq (vi) n (vii) 0 and 1 (viii) random numbers
(ix) independent (x) varies/changes

3.3 (i) b (ii) c (iii) c (iv) d (v) b
(vi) c (vii) d (viii) a (ix) b (x) d

3.10 (i) 0 (ii) 0.29 (iii) 0 (iv) 0.936 (v) 0.352

3.11

X	0	1	2	3	4	5
p(x)	0.095	0.286	0.343	0.206	0.062	0.007

3.12 (i) 0.656 (ii) 0.891

3.13 (i) 0.07776 (ii) 0.2592 (iii) 0.92224

3.14 (i) 0.60501 (ii) 0.39499

3.15 0.227

3.16 0.62343

3.17 0.0005

3.19 (i) 40 bolts (ii) 6

3.20 $n = 12$, $p = 0.25$

3.21 $n = 41$, $p = 0.302$

3.22 No, because $q = 1.8$ which cannot be greater than 1.

3.23 (i) 16 (ii) 0.9576

3.24 mean = 3, variance = 1.6, S.D = 1.25

3.25 (b) 52.08, 41.67, 12.5, 1.67, 0.08

3.26 mean = 5.42 variance = 2.4824

3.27 11.04 64.07 154.94 199.83 144.98 56.10 9.04

3.30

x	1	2	3	4	5
p(x)	1/66	2/11	5/11	10/33	1/22

Mean = 3.187, Variance = 0.706

3.31 0.275

3.32 0.213

Exercise-4

4.1 (i) F (ii) F (iii) T (iv) T (v) T
(vi) T (vii) F (viii) T (ix) F (x) T

4.2 (i) 4 (ii) $Z \sim N(0,1)$ (iii) $X = \mu$ (iv) $\frac{4}{5}\sigma$
(v) zero (vi) normal (vii) 10 (viii) mean = μ
(ix) Mean = 0 (x) 95.4

4.3 (i) b (ii) a (iii) b (iv) b (v) c
(vi) c (vii) c (viii) b (ix) d (x) a

4.8 (i) 0.0548 (ii) 0.044 (iii) 0.103 (iv) 0.475 (v) 0.950
(vi) 0.682 (vii) 0.006

4.9 (i) 0.5987 (ii) 0.2119 (iii) 0.0668 (iv) 0.7734

4.10 (i) 0.0043 (ii) 0.0735 (iii) 0.0103 (iv) 0.1196 (v) 0.9777
(vi) 0.9850

4.11 (i) z = 2.58 (ii) z = 1.24 (iii) z = −1.75

4.12 (i) z = 0.84 (ii) z = −0.25 (iii) z = −1.04 and +1.04

4.13 (i) 0.1587 (ii) 0.0228 (iii) 0.3721

4.14 (i) 0.9452 (ii) 0.0548 (iii) 0.3085 (iv) 0.6915 (v) 0.6898
(vi) 0.1832

4.15 0.1359

4.16 0.2515

4.17 (i) 0.9798 = 97.98% (ii) 0.3531 = 35.31%
(iii) 0.0559 = 5.59%

4.18 (ii) 0.0548 (ii) 0.4772

4.19 (i) 0.3446, 173 (ii) 0.3811, 191

4.20 (i) 758 (ii) 629 (iii) 206

4.21 0.9554 = 95.54%

4.22 0.1587, 159

4.23 (i) 0.6826 (ii) 0.0317

4.24 0.4706

4.25 (i) 0.0167 (ii) 0.3885 (iii) 0.0775

4.26 (i) 92.22 (ii) 4.44

4.27 M.D = 4

4.28 $\sigma = 20$

Exercise-5

5.1 (i) T (ii) T (iii) F (iv) F (v) F
(vi) T (vii) F (viii) T (ix) T (x) F

5.2
(i) finite (ii) infinite (iii) census (iv) sampling error
(v) standard error (vi) sampling frame (vii) more than once
(viii) sampling fraction (ix) fpc (x) $n < 5\%N$

5.3 (i) a (ii) b (iii) b (iv) d (v) d
(vi) c (vii) c (viii) c (ix) d (x) d

5.16 (b) $n_1=8$ $n_2=9$ $n_3=11$ $n_4=12$

5.17 $n_1=75$ $n_2=50$ $n_3=25$ $n_4=40$ $n_5=10$

5.19 (b) $\mu_{\overline{X}} = 6, \sigma_{\overline{X}} = 1$

5.20 $\mu_{\overline{X}} = 3.8, \quad \sigma_{\overline{X}} = 0.98$

5.21 $\mu_{\overline{X}} = 50 \quad \sigma_{\overline{X}} = 6$

5.22 (i) z =1 (ii) z =2

5.23 (i) $\mu = 7, \ \sigma^2 = 5.2$ (ii) 0.87 (iii) 1.73

5.24 (i) $\mu_{\overline{X}} = \mu = 9$ (ii) $\sigma = 6.48, \quad \sigma_{\overline{X}} = 2.9$

5.26 $\mu_{\overline{X}_1-\overline{X}_2} = 1 \qquad \sigma_{\overline{X}_1-\overline{X}_2} = 1.732$

5.27 (b) $\mu_{\overline{X}_1-\overline{X}_2} = 2 \quad \sigma^2_{\overline{X}_1-\overline{X}_2} = 2.1667$

5.29 $\mu_{\hat{p}} = 0.5, \quad \sigma^2_{\hat{p}} = 0.05$

5.30 $\hat{p} = \frac{3}{7} \qquad \mu_{\hat{p}} = \frac{3}{7} \qquad \sigma^2_{\hat{p}} = \frac{5}{49}$

5.31 $E(\hat{p}_1 - \hat{p}_2) = 0 \ V(\hat{p}_1 - \hat{p}_2) = \frac{1}{9}$

Exercise-6

6.1 (i) F (ii) T (iii) F (iv) F (v) T
(vi) F (vii) T (viii) T (ix) T (x) F

6.2 (i) random variable (ii) estimate (iii) estimator
(iv) point (v) unbiased (vi) efficient
(vii) shortest (viii) increasing (ix) two
(x) confidence coefficient

6.3 (i) b (ii) c (iii) b (iv) c (v) a
(vi) d (vii) d (viii) b (ix) c (x) a

6.9 (i) $\overline{X}$ =5.9 (ii) $s^2 = 5.8178$ (iii) $s_{\overline{X}} = 0.7667$

6.10 (i) $E(\overline{X}) = \mu = 14$ (ii) $E(s^2) = \sigma^2 = 9.333$

6.14 ($9.84 < \mu < 12.16$)

6.15 ($0.774 < \mu < 0.826$)

6.16 ($788.24 < \mu < 811.76$)

6.17 ($-2.27 < \mu < 2.67$)

6.18 ($63.412 < \mu < 64.588$)

6.19 ($31.59 < \mu < 54.93$)

6.20 ($2.4 < \mu_1 - \mu_2 < 7.6$)

6.21 ($0.55 < \mu_2 - \mu_1 < 1.25$)

6.23 ($4.3371 < \mu < 4.4229$)

6.24 ($33.52 < \mu < 35.28$)

6.25 ($-15.3 < \mu_A - \mu_B < 1.3$)

6.26 ($0.45 < P < 0.75$)

6.27 ($0.62 < P < 0.74$)

6.28 ($-0.206 < P_1 - P_2 < -0.120$)

Exercise-7

7.1 (i) F (ii) T (iii) F (iv) F (v) T
(vi) T (vii) T (viii) F (ix) T (x) F

7.2 (i) assertion (ii) null (iii) alternative
(iv) composite (v) first, second (vi) level of significance
(vii) 0, 1 (viii) rejection region (ix) rule
(x) degrees of freedom

7.3 (i) c (ii) a (iii) b (iv) a (v) c
(vi) a (vii) c (viii) b (ix) b (x) b

7.10 $z = 2.61$, reject H_0

7.11 $z = 2.4$, reject H_0

7.12 $z = -2.88$, reject H_0

7.13 $z = 2.33$, reject H_0

7.14 $z = -1.93$, reject H_0

7.16 $t = 0.316$, accept H_0

7.17 $t = 1.844$, accept H_0

7.19 $z = 4.22$, reject H_0

7.20 $z = -1.334$, accept H_0

7.22 $t = 3.33$, reject H_0

7.23 $t = 3.05$, reject H_0

7.25 $z = 2.19$, reject H_0

7.26 $z = -0.73$, accept H_0

7.27 $z = -2.1004$, accept H_0

7.29 $z = -3.1623$, reject H_0

7.30 $z = 9.6825$, reject H_0

Exercise-8

8.1 (i) F (ii) T (iii) T (iv) F (v) T
(vi) F (vii) T (viii) F (ix) T (x) F

8.2 (i) dichotomous (ii) order (iii) second (iv) zero
(v) $\frac{(A)(B)}{n}$ (vi) zero (vii) not same
(viii) -1 *to* $+1$ (ix) equal (x) 20

8.3 (i) c (ii) b (iii) d (iv) d (v) b
(vi) c (vii) b (viii) c (ix) c (x) b

8.9 $Q = -0.76$, the association between extravagant fathers and extravagant sons is negative.

8.10 $Q = 0$, attributes A and B are independent.

8.11 $Q = 0.46$, there is positive association.

8.12 $Q = 0.9$, there is high degree of positive association between intelligent fathers and sons.

8.13 $Q = -0.57$, it means that vaccine and small-pox are negatively associated i.e. vaccine prevents the attack of small-pox.

8.14 $\chi^2 = 48.24$, reject H_0

8.15 $\chi^2 = 8.89$, reject H_0

8.16 $\chi^2 = 60.183$, reject H_0

8.18 $\chi^2 = 598.51$, reject H_0

8.19 $\chi^2 = 29.79$, reject H_0

8.20 $\chi^2 = 763.76$, reject H_0

8.22 $\chi^2 = 3.279$, C = 0.067

8.23 $\chi^2 = 1048.5$, C = 0.3681

Exercise-9

9.1 (i) T (ii) T (iii) F (iv) T (v) T
(vi) T (vii) T (viii) F (ix) T (x) F

9.2 (i) experimental unit (ii) randomization (iii) replication
(iv) extraneous factors (v) efficient
(vi) component variances (vii) no problem
(viii) can never (ix) replication
(x) field experiments

9.3 (i) c (ii) c (iii) a (iv) b (v) b
(vi) a (vii) d (viii) c (ix) d (x) d

9.16 $F_c = 0.277$ accept H_0

9.17 $F_c = 1.37$ accept H_0

9.18 $F_c = 0.055$ accept H_0

9.19 $F_c = 4$ reject H_0

9.20 $F_c = 3.333$ accept H_0

9.21 $F_c = 5.56$ reject H_0

9.22 $F_c = 2.56$ accept H_0

9.23 $F_c = 0.82$ accept H_0

9.24 $F_c = 2.10$ accept H_0

9.25 $F_c = 6.46$ accept H_0

Glossary

Addition rule: Rule for determining the probability that, on a single trial, either event *A* occurs, or event *B* occurs, or they both occur.

Alternative hypothesis: Statement that is equivalent to the negation of the null hypothesis.

Analysis of variance: Method of analysing population variances in order to test hypotheses about means of populations.

Binomial experiment: Experiment with a fixed number of independent trials, where each outcome falls into exactly one of two categories.

Block: A group of subjects that is similar in the ways that might affect the outcome of an experiment.

Census: Collection of data from every element in a population.

Central limit theorem: Theorem stating that sample means tend to be normally distributed.

Chi-square distribution: A continuous probability distribution.

Classical approach to probability: Approach in which the probability of an event is determined by dividing the number of ways the event can occur by the total number of possible outcomes.

Combinations rule: Rule for determining the number of different combinations of selected items.

Complement of an event: All outcomes in which the original event does not occur.

Completely randomized design: An experiment where by each element is given the same chance of belonging to the different categories or treatments.

Compound event: Combination of simple events.

Conditional probability: The probability of an event, given that some other event has already occurred.

Confidence coefficient: Probability that a population parameter is contained within a particular confidence interval; also called confidence level or degree of confidence.

Confidence interval limits: Two numbers that are used as the high and low boundaries of a confidence interval.

Confidence interval: Range of values used to estimate some population parameter with a specific confidence level; also called an interval estimate.

Confidence level: Probability that a population parameter is contained within a particular confidence interval.

Contingency table: Table of observed frequencies where the rows correspond to one variable of classification and the columns correspond to another variable of classification; also called a two-way table.

Continuity correction: Adjustment made when a discrete random variable is being approximated by a continuous random variable.

Continuous data: Data resulting from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

Continuous random variable: A random variable with infinite values that can be associated with points on a continuous line interval.

Critical region: The set of all values of the test statistic that would cause rejection of the null hypothesis.

Critical value: Value separating the critical region from the values of the test statistic that would not lead to rejection of the null hypothesis.

Degree of confidence: Probability that a population parameter is contained within a particular confidence interval; also called level of confidence.

Degrees of freedom: Number of values that are free to vary after certain restrictions has been imposed on all values.

Dependent events: Events for which the occurrence of any one event affects the probabilities of the occurrences of the other events.

Dependent sample: Sample whose values are related to the values in another sample.

Discrete random variable: Random variable with either a finite number of values or a countable number of values.

Disjoint events: Events that cannot occur simultaneously.

Efficiency: It is a criterion for selection of an efficient estimator.

Estimate: Specific value or range of values used to approximate some population parameter.

Estimator: Sample statistic (such as the sample mean) used to approximate a population parameter.

Event: The collection of favorable outcomes to a happening from the sample space.

Expected value: For a discrete random variable, the mean value of the outcomes

Experiment: Application of some treatment followed by observation of its effects on the subjects.

Experimental units: Subjects in an experiment.

Factorial rule: Rule stating that n different items can be arranged $n!$ different ways.

Finite population correction factor: Factor for correcting the standard error of the mean when a sample size exceeds 5% of the size of a finite population.

Fundamental counting rule: Rule stating that, for a sequence of two events in which the first event can occurs m ways and the second can occur n ways, the events together can occur a total of mn ways.

Hypothesis: Statement or claim about some property of a population.

Hypothesis test: Method for testing claims made about populations; also called test of significance.

Independent events: Events for which the occurrence of any one of the events does not affect the probabilities of the occurrences of the other events.

Inferential statistics: Collection of methods that help make decisions about a population based on sample results.

Interval estimate: Range of values used to estimate some population parameter with a specific level of confidence; also called a confidence interval.

Left-tailed test: Hypothesis test in which the critical region is located in the extreme left area of the probability distribution.

Level of confidence: Probability that a population parameter is contained within a particular confidence interval; also called degree of confidence.

Multiplication rule: Rule for determining the probability that event A will occur on one trial and event B will occur on a second trial.

Mutually exclusive events: Events that cannot occur simultaneously.

Non sampling errors: Errors from external factors not related to sampling.

Null hypothesis: Claim made about some population characteristic, usually involving the case of no difference.

Odds against: Ratio of the probability of an event not occurring to the event occurring, usually expressed in the form of a: b where a and b are integers having no common factors.

Odds in favour: Ratio of the probability of an event occurring to the event not occurring, usually expressed as the ratio of two integers with no common factors.

One-way analysis of variance: Analysis of variance involving data classified into groups according to a single criterion only.

Paired samples: Two sample which are dependent in the sense, that the data values are matched by pair.

Parameter: A summary measure calculated for population data.

Point estimate: Single value that serves as an estimate of a population parameter.

Pooled estimate of variance: Estimate of the variance that is common to two populations, found by computing a weighted average of the two sample variances.

Population or target population: The collection of all elements whose characteristics are being studied.

Probability distribution: Collection of values of a random variable along with their corresponding probabilities.

Probability: Measure of the likelihood that a given event will occur expressed as a number between 0 and 1.

Qualitative or categorical variable: A variable that cannot assume numerical values but is classified into two or more categories.

Random sample: Sample selected in a way that allows every member of the population to have the same chance of being chosen.

Random selection: Selection of sample elements in such a way that all elements available for selection have the same chance of being selected.

Random variable: Variable (typically represented by X) that has a single numerical value (determined by chance) for each outcome of an experiment.

Representative sample: A sample that contains the same characteristics as the corresponding population.

Right-tailed test: Hypothesis test in which the critical region is located in the extreme right area of the probability distribution.

Sample: A portion of the population of interest.

Sample size: Number of items in a sample.

Sample space: Set of all possible outcomes or events in an experiment that cannot be further broken down.

Sample survey: A survey that includes elements of a sample.

Sampling distribution of proportion: The probability distribution of sample proportions, with all samples having the same sample size n.

Sampling distribution of sample means: Distribution of the sample means that is obtained when we repeatedly draw samples of the same size from the same population.

Sampling error: Difference between a sample result and the true population result; results from chance sample fluctuations.

Sampling variability: Variations of a statistic in different samples.

Significance level: Probability of making a type I error when conducting a hypothesis test.

Simple event: Experimental outcome that cannot be further broken down.

Simple random sample: Sample of a particular size selected so that every possible sample of the same size has the same chance of being chosen.

Standard normal distribution: Normal distribution with a mean of 0 and a standard deviation equal to 1

Standard score: Number of standard deviations that a given value is above or below the mean; also called z-score.

Statistic: A summary measure calculated for sample data.

Stratified sampling: The sampling method in which samples are drawn from each stratum.

Subjective probability: Guess or estimate of a probability based on knowledge of relevant circumstances.

Survey: Collection of data on the elements of a population or sample.

Systematic sampling: Sampling in which every k^{th} element is selected.

t-distribution: Bell-shaped distribution usually associated with sample data from a population with an unknown standard deviation.

Test of independence: Test of the null hypothesis that for a contingency table, the row variable and column variable are not related.

Test statistic: Sample statistic based on the sample data; used in making the decision about rejection of the null hypothesis.

Treatment: Property or characteristic that allows us to distinguish the different populations from one another; used in analysis of variance.

Tree diagram: Graphical depiction of the different possible outcomes in a compound event.

Two-tailed test: Hypothesis test in which the critical region is divided between the left and right extreme areas of the probability distribution.

Type I error: Rejecting the null hypothesis when it is true.

Type II error: Accept the null hypothesis when it is false.

Unbiased estimator: Sample statistic that tends to target the population parameter that it is used to estimate.

Uniform distribution: Probability distribution in which every value of the random variable is equally likely.

Variable: A characteristic under study or investigation that assumes different values for different elements.

Variance between samples: In analysis of variance, the variation among the different samples.

Variation within samples: In analysis of variance, the variation that is due to chance.

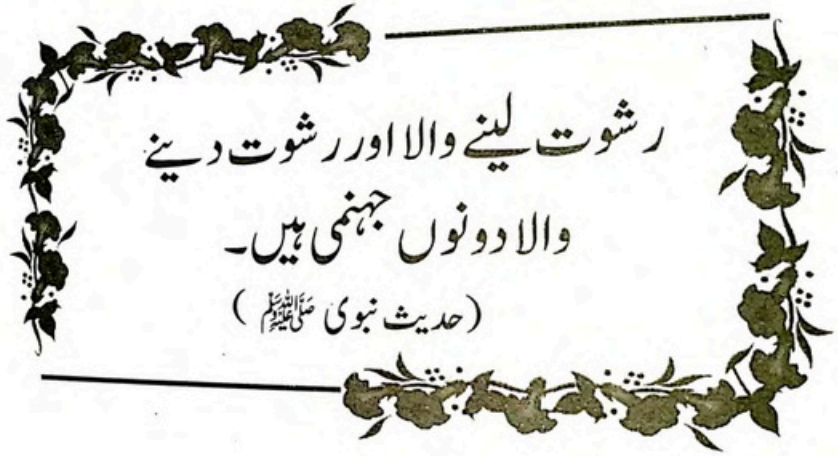
Z-score: Number of standard deviations that a given value is above or below the mean.

INDEX

About the Author

Jamal Shah is Professor of Statistics. He was born in 1964 in Mangah, District Mardan. He got his early education from Government higher secondary school Dargai Mangah district Charsadda. He did F.Sc. from Government postgraduate college Mardan, B.Sc. from Government postgraduate Jehanzeb College Saidu Sharif Swat. In 1988, from university of Peshawar, he received his M.Sc. Statistics degree and stood third in his batch.

He started his service in December, 1988 as Lecturer in statistics from Government Khushal Khan Khattak college Akora Khattak district Nowshera. He remained chairman department of Statistics, Government postgraduate college Mansehra and Abbottabad for many years. Presently he is working as Principal, Govt. Degree college, Zaida Swabi. The author has a vast teaching experience of 30 years in the subject statistics at Inter, Bachelor and Master level.



قومی ترانہ

پاک سر زمین شاد باد　　کشورِ حسین شاد باد
تو نِشانِ عزمِ عالی شان　　ارضِ　　پاکستان
مرکزِ یقین شاد باد

پاک سَر زمین کا نِظام　　قُوّتِ اُخوّتِ عوام
قوم، مُلک، سلطنت　　پائندہ تابندہ بَاد
شاد باد منزلِ مُراد

پرچمِ ستارۂ و ہلال　　رَہبرِ ترقی و کمال
ترجمان ماضی شانِ حال　　جانِ　　اِستقبال
سایۂ خُدائے ذُوالجلال

Printed by
Etihad

Code No.

Quabtity