PREFACE

In the curricular structure introduced by this University for students of Post Graduate degree programme, the opportunity to pursue Post-Graduate course in a subject is introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with, the objectives of open, education which does not believe in artificial differentiation. I am happy to note that the university has'been recently accredited by National Assessment and Accreditation Council of India (NAAC) with Ififf grade 'A'.

Keeping this in view, the study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars. 1 is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing, and devising of a proper layout of the materials. Practically speaking, their role amounts to an involvement in invisible teaching'. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by Kythe other.

The more a learner would seriously pursue these study materials, the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that they may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great deal of these efforts is still experimental-to fact, pioneering in certain areas. Naturally, there is eveiy possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Subha Sankar Sarkar
Vice-Chancellor

Netaji Subhas Open University

Post Graduate Degree Programme

Subject : Commerce (M.Com)

Course : Statistics for Managerial Decisions

Course Code : PGCO-VII

First Print : May, 2022

Printed in accordance with the regulations of the Distance Education Bureau of the University Grants Commission.

Netaji Subhas Open University
Post Graduate Degree Programme
Subject : Commerce (M.Com)
Course : Statistics for Managerial Decisions
Course Code : PGCO-VII

Netaji Subhas Open University

PG : Commerce
(M. Com)
(New Syllabus)

Course : Statistics for Business Decision

Course Code : PGCO-VII

Module - I

Module - II

Unit 1 ❑ Probability Theory

Structure

1.0 Objectives

After studying the present unit, you will be able to (i) understand the basic concept of probability (ii) discuss the different terms associated with the probability theory and (iii) explain the various theorems of probability

1.1 Introduction

The understanding of probability becomes essential for decision makers because the occurrence and non-occurrence of certain events are vital for decision making in business and other spheres of life. The term 'probability' implies chance. The genesis of probability lies in the concept of gambling. The development of probability theory can be attributed to a gambler's dispute in France in 1654. The problem was reported to two famous mathematicians in France, Blaise Pascal and Pierre de Fermat. This has resulted in an exchange of letters between these two mathematicians in which the

basic principles of probability were designed for the first time. In the subsequent years famous mathematicians like Laplace, Bernoulli, Markov etc. have made notable contribution towards developing the theory of probability.

1.2 Basic Concept of Probability

The literal meaning of the term 'Probability' is ‘chance’. Its theory deals with these experiments (Random experiments) whose results depend on chance and cannot be predicted with full certainty in advance.

There are two different contents in which the word probability may be used : (a) in regard to some proposition. (b) regarding the outcomes of an experiment that can, at least conceivably, be repeated infinite number of times under essentially similar conditions. In statistics, the theory mainly deals with the second context.

1.3 Some Important Terms and Definitions

Random experiment : Random experiment is an experiment or an act which can be repeated under indentical conditions, such that its results (outcomes) depend on chance and nobody can assert with certainty which result will occur (materialise) in any particular trial. However, all the possible outcomes are known in advance. Some examples are (a) Drawing cards from a pack, (b) Throwing a die, (c) Counting a number of cars on a gives road crossing during a specific time period of a day. (d) Counting the number of male students (or female students) present in a class on a specific day. etc.

Event : The outcomes of a random experiment are called events. It is any phenomenon that can occur in a random experiment. Events (Generally denoted by capital letters like A, B, C, A_1, A_2, B_1 etc.) can be 'elementary' events or 'composite' events.

Elementary events : Elementary events of a random experiment are those events which can not be further decomposed into simpler events. e.g. In the throwing of a sixed-faced die the occurence of any of the face number points 1, 2, 3, 4, 5, 6 are elementary events.

Composite events : A composite event is an aggregate of several elementary events e.g. In the throwing of a die, if the event 'Face number points which are even numbers' (i.e. the event which is an aggregate of three elementary events, face number points 2, 4, 6) is considered, then it is a composite event.

Sample Space : The whole set of elementary events of a random experiment is called the sample space of that experiment. It is also called the sure event.

Mutually Exclusive Events : Events are said to be mutually exculsive if two or more of them can not occur simultaneously. i.e. The occurence of any one of those events in a particular tiral, ifself confirms the non-occurence of the remaining events. e.g. if 'A' denotes the event that even number appears on the upper most face of the die, and event 'B' denotes that odd number appears on the upper most face of the die, then the events 'A' and 'B' are mutually exclusive.

Exhaustive Events : Several set of events are said to form on exhaustive set of events, if at least one of them must necessarily occur in any trial of the random experiment. e.g. in the throwing of a die, the elementary events with face number points 1, 2, 3, 4, 4 and 6 together form an exhaustive set.

Trial : Any particular performance of the random experiment is called a trial.

Cases Favourable to an Event : Out of all the possible events of a random experiment, those cases which entail the occurence of an event A are called 'cases favourable to A'.

Equally Likely Events : The outcomes of a random experiment are said to be 'equally likely', if after taking into consideration all relevant evidences, none of them can be expected in preference to another. e.g. If a card is drawn from a full pack of 52 cards, the outcomes are equally likely.

Union of events : If there are two events 'A' and 'B', then by their union, i.e. $A \cup B$, we mean the occurence of either A or B or both A and B. similarly, for multiple events A_1, A_2, A_3,...A_n, their union is given by $\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup A_3 \cup ... \cup A_n$. It denotes the occurence of at least one of the events A_1, A_2,...A_n.

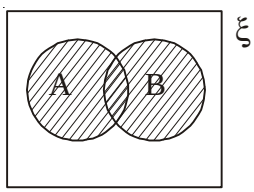


Fig 1 : $A \cup B$ is represented by the shaded area.

Intersection of events : By the intersection of two events, A and B, i.e. $A \cap B$,

we mean the occurence of both A and B. Similarly, for multiple events A_1, A_2, A_3...A_n, their intersection is given by $\bigcap_{i=1}^{n} A_i = A_1 \bigcap A_2 \bigcap A_3 \bigcap ... \bigcap A_n$. It denotes the simultaneous occurence of A_1, A_2, A_3,...A_n.

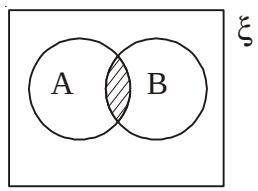


Fig 2 : A⋂B is the shaded area.

Difference of Events : A-B will denote the occurence of A together with the non-occurence of B.

$A - B = A - A \bigcap B$

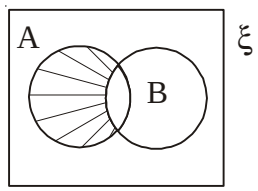


Fig 3 : A–B is the shaded area.

Complement : By the complement of an event A, i.e. A^C or $\bar{A}$ or A′ we mean the non occurrence of event A.

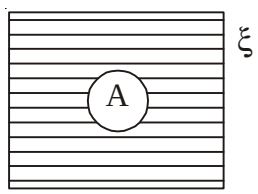


Fig 4 : The shaded area is A^C

In all the above Venn diagrams, the collection of geometric points within a given

region is taken to represent the sample space and the collection of points within an oval/circle is taken to represent an event.

Some properties of the operations of forming union and intersection of the events :

Commutativity : $A \cup B = B \cup A$, $A \cap B = B \cap A$.

Associativity : $A \cup (B \cup C) = (A \cup B) \cup C$,

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

Idempotency : $A \cup A = A$, $A \cap A = A$

De Morgan's rules : $(A^C)^C = A$, $(A \cup B)^C = A^C \cap B^C$

$(A \cap B)^C = A^C \cup B^C$

Classicial Definition of Probability :

In classical approach, for the sake of simplicity, we shall assume that the total number of elementary events in the sample space is finite. We shall further assume that the elementary events are equally likely.

Thus if the total number of elementary events in the sample space is M and out of those events, M(A) number of elementary events are favourable to the event A (such that A happens if and only if one of these elementary events happens), then the probability of the event A is given by $P(A) = \frac{M(A)}{M}$

$$\left[\text{i.e } P(A) = \frac{\begin{array}{c}\text{(Number of elementary events favourable}\\ \text{to the event A)}\end{array}}{\begin{array}{c}\text{(Total number of mutually exclusive, exhaustiv and}\\ \text{equally likely elementary events in the sample space)}\end{array}}\right]$$

The above equation gives the classical definition of probability, and it is applicable when the sample space is finite and the elementary events are equally likely.

From the classical definition it follows that

(a) If A is an impossible event, then M (A) = 0. Hence P(A) = 0

(b) If A is a sure event, then M(A) = M. Hence P(A) = 1.

(c) For any event A, we have $0 \leq M(A) \leq M$

or, $\frac{0}{M} \leq \frac{M(A)}{M} \leq \frac{M}{M}$; i.e. $0 \leq P(A) \leq 1$

(d) If the occurence of event A implies the occurence of event B, then every elementary event that is favourable to A is also favourable to B. So $M(A) \leq M(B)$;

or, $\frac{M(A)}{M} \leq \frac{M(B)}{M}$

or, $P(A) \leq P(B)$

Illustration 1.1 : If a sixed faced die is thrown, what is the probability that the number appearing uppermost is at least equal to 5?

Solution : Regarding the number appearing on the uppermost face of a die, there are six possible cases, viz. 1, 2, 3, 4, 5 and 6. So the sample space is defined by these six elementary events. Now, if the die is regular in shape and is made of homogenous material and if during the throw no preference is given to any specific side, then the elementary events may be considered equally likely. Out of the six cases, there are two cases which are favourable to the event 'at least equal to 5'. viz. 5 and 6.

So the required probability is $\frac{2}{6} = \frac{1}{3}$

Illustration 1.2 : If three coins are tossed simultaneously, what is the probability of getting three heads?

Solution : The tossing of each coin may result in a head {H} or a tail {T}. So, when three coins are tossed simultaneously, the sample space can be written as—

$\xi = \{HHH, HHT, HTH, THH, HTT, TTH, THT, TTT\}$

i.e. There are totally 8 elementary events in the sample space.

Now, if the event of getting 3 heads is denoted by A, there A = {HHH}. i.e. the number of elementary events favourable to event A is 1. So the required probability is $P(A) = \frac{1}{8}$.

Illustration 1.3 : The digits 1, 2, 3, 4 and 5 are written down in random order

to give a 5-digit number. what is the probability that the number is divisible by either 2 or 5?

Solution : Totally, the digits can be arranged in $\lfloor\underline{5}$ ways. i.e. the sample space is made of $\lfloor\underline{5}$ number of elementary events. The fact that the digits are placed in random order implies that the $\lfloor\underline{5}$ ways (i.e. the elementary events) are to be considered equally likely. Now, if the number is divisible by either 2 or 5, then the last (i.e. the 5th) digit of the number should be any of the following 3 digits : 2, 4, 5. In each of those three cases, the first 4 places of the number can be filled in $\lfloor\underline{4}$ ways. So, if 'A' denotes the event that 'the number is divisible by either 2 or 5', then the total number of cases favourable to probability is given by $P(A)=\frac{3\times\lfloor\underline{4}}{\lfloor\underline{5}}=\frac{3}{5}$.

Illustration 1.4 : A box contians 5 white and 7 black balls. One ball is drawn at random from the box. What is the probability that it is black?

Solution : Totally, the box contains 12 balls. So, if a ball is drawn at random from the box, then there are ${}^{12}C_1 = 12$ possible outcomes (which are equally likely, mutually exclusive and exhaustive). Out of the 7 black balls contained in the box, one black ball can be drawn at random in ${}^{7}C_1 = 7$ ways. So, if A denotes the event that the 'drawn ball is black', then the number of cases favourable to the event A is 7.

So, the required probability as given by $P(A)=\frac{7}{12}$

Illustration 1.5 : There are 7 distinguishable balls which are to be distributed at random into 3 boxes. What is the probability that a specified box contains exactly 3 balls?

Solution : The first ball may be placed in any of the three boxes and hence it may be distributed in 3 ways. The second ball may also be distributed in 3 ways, the third ball may also be distributed in 3 ways and so on and finally the 7th ball may also be distributed in 3 ways. So, the total number of ways in whichs the 7 balls may be distributed in the 3 boxes is $3 \times 3 \times 3 \times$...(Items) = 3^7.

Now, there are ${}^{7}C_3$ number of possible combinations of 3 balls that can be made of the available 7 balls. Any one of these ${}^{7}C_3$ possible combination (of 3 balls) may be placed in the specified box which will contain exactly 3 balls. Once, the three balls

have been placed in the specified box, the remaining 4 balls may be placed in the remaining two boxes in 2^4 ways. So, the total number of favourable cases is ${}^7C_3 \times 2^4 = 35 \times 2^4$.

So the required probability is $\frac{35\times 2^4}{3^7}$

1.4 Theorems on Probability

1.4.1 Theorem of Total Probability : (Addition Theorem)

If two events A_1 and A_2 are mutually exclusive, then the probability of occurrence of either A_1 or A_2 is given by $P(A_1 \cup A_2) = P(A_1) + P(A_2)$

Proof : Let 'M' be the total number of elementary event in the sample space. Out of these M events, $M(A_1)$ number of elementary events are favourable to the event A_1 and $M(A_2)$ number of elementary events are favourable to the event A_2.

Due to the fact that A_1 and A_2 are mutually exclusive events, the number of elementary events favourable to either A_1 or A_2 is $M(A_1 \cup A_2) = M(A_1) + M(A_2)$

So, $$P(A_1 \cup A_2) = \frac{M(A_1)+M(A_2)}{M}$$
$$= \frac{M(A_1)}{M} + \frac{M(A_2)}{M}$$
$$= P(A_1) + P(A_2)$$

This rule can be extended for n (>2) mutually exclusive events $A_1, A_2, A_3 ... A_n$ and one can write

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

or, $$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

Some Deductions from the Theorem of Total Probability :

(a) If A^C denotes the event complementary to the event A, then

$P(A) + P(A^C) = 1$ {Since A and A^c are mutually exclusive and exhaustive events}

or, $P(A^C) = 1 - P(A)$.

(b) For any two events A_1 and A_2

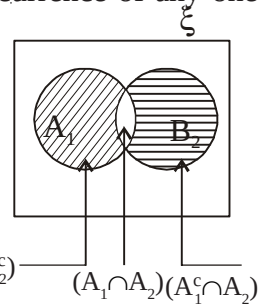
$P(A_1 - A_2) = P(A_1) - P(A_1 \cap A_2)$

If A_1 and A_2 are mutually exclusive events, theorem $P(A_1 \cap A_2) = 0$ and then, $P(A_1 - A_2) = P(A_1)$.

(c) When events A_1 and A_2 are not necessarily mutually exclusive, then

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

Proof : the occurence of the event $(A_1 \cup A_2)$ implies the occurrence of any one of the mutually exclusive events.



$(A_1 \cap A_2^C), (A_1 \cap A_2)$ and $(A_1^C \cap A_2)$

So, $P(A_1 \cup A_2) = P\left[(A_1 \cap A_2^C) \cup (A_1 \cap A_2) \cup (A_1^C \cap A_2)\right]$

$= P(A_1 \cap A_2^C) + P(A_1 \cap A_2) + P(A_1^C \cap A_2)$

$= \left[P(A_1) - P(A_1 \cap A_2)\right] + P(A_1 \cap A_2) + \left[P(A_2) - P(A_1 \cap A_2)\right]$

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

$\because P(A_1) = P(A_1 \cap A_2^C) + P(A_2 \cap A_2)$,
$P(A_1 \cap A_2^C) = P(A_1) - P(A_1 \cap A_2)$
$\because P(A_2) = P(A_1^C \cap A_2) + P(A_1 \cap A_2)$,
$P(A_1^C \cap A_2) = P(A_2) - P(A_1 \cap A_2)$

Similarly, if there are three events, A_1, A_2 and A_3 which are not necessarily mutually exclusive, then

$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) -$

$P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$

Proof : For any two events, B and C, it has been proved that $P(B \cup C) = P(B) + P(C) - P(B \cap C)$

So, by induction,

$P\left[A_1 \cup A_2 \cup A_3\right] = P\left[(A_1 \cup A_2) \cup A_3\right]$

$= P(A_1 \cup A_2) + P(A_3) - P[(A_1 \cup A_2) \cap A_3]$

$= P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) - P[(A_1 \cap A_3) \cap (A_2 \cap A_3)]$

$= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - [P(A_1 \cap A_3) + P(A_2 \cap A_3) - P(A_1 \cap A_2 \cap A_3)]$

$= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$

[**Note :** This rule can be extended for 'n' (>3) number of events in the following manner :

Theorem : Whatever be the events $A_1, A_2, ...A_n$,

$P(A_1 \cup A_2 \cup ... \cup A_n) = S_1 - S_2 + S_3 ... + (-1)^{n-1} S_n$

Where $S_1 = \sum_{i=1}^{n} P(A_i)$, $S_2 = \sum_{\substack{i,j=1 \\ i<j}}^{n} P(A_i \cap A_j)$, $S_3 = \sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} P(A_i \cap A_j \cap A_k)$ and so on]

(d) For any two events A_1 and A_2,

$P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$ [Boole's inequality]

Proof : For any two events A_1 and A_2, which are not necessarily mutually exclusive,

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

Now $P(A_1 A_2) \geq 0$. Being a probability. $P(A_1 \cap A_2)$ can never be negative, and its minimum value is 0. (which occurs when events A_1 and A_2 are mutually excclusive).

So, $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$

Illustration 1.6 : Three lots contain respectively 15%, 10% and 20% defective items. If one items is drawn at random from each lot, then what is the probability that among the three items drawn, at least one item is defective?

Solution : Let A_1 be the event of drawing a defective item from the lot having 15% defective items, A_2 be the event of drawing a defective item from the lot having 10% defective items and A_3 be the event of drawing a defective item from the lot having 20% defective items.

Then $P(A_1) = 0.15$; and $P(A_1^C) = 1 - P(A_1) = 0{\cdot}85$

$P(A_2) = 0.10$; and $P(A_2^C) = 1 - P(A_2) = 0{\cdot}9$

$P(A_3) = 0.20$; and $P(A_3^C) = 1 - P(A_3) = 0{\cdot}80$.

So the probability of drawing at least one defective item = $P(A_1 \cup A_2 \cup A_3)$

$= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$

$= 0{\cdot}15 + 0{\cdot}10 + 0{\cdot}20 - (0{\cdot}15) \times (0{\cdot}10) - (0{\cdot}15) \times (0{\cdot}20) - (0{\cdot}10) \times (0{\cdot}20) + (0{\cdot}15) \times (0{\cdot}10) \times (0{\cdot}20)$

$= 0{\cdot}45 + [0{\cdot}015 + 0{\cdot}03 + 0{\cdot}02] + 0{\cdot}003$

$= 0.453\ 0.065 = 0.388$

The same result may also be obtained in the following way :

The probability of drawing at least one defective item = 1 – The probability of drawing all three non-defective items.

$= 1 - P(A_1^C \cap A_2^C \cap A_3^C)$

$= 1 - P(A_1^C) \cdot P(A_2^C) \cdot P(A_3^C)$

$= 1 - (0.85) \times (0.9) \times (0.80) = 1 - 0.612 = 0.388.$

Illustration 1.7 : The probability that a student will pass an accountancy examination is $\frac{3}{4}$, and the probability that he will pass a statistic examinations is $\frac{5}{9}$. If the probability of the student passing at least one of the two subjects is $\frac{5}{6}$, then what is the probability that he will pass both the subjects examinations?

Solution : Let

A_1 denotes that the student will pass the accountancy examination

A_2 ,, ,, ,, ,, ,, ,, ,, statistics examination

Then, $P(A_1)\frac{3}{4}$; $P(A_2)\frac{5}{9}$

Also, $P(A_1 \cup A_2)$ = The probability of the student passing at least one of the two subjects $= \frac{5}{6}$

We know that $P(A_1 \cup A_2) = P(A_1) + P(A_2) - (A_1 \cup A_2)$

$$= \frac{3}{4} + \frac{5}{9} - \frac{5}{6} = \frac{17}{36}$$

So, the probability that the student will pass the examinations of both the subjects is $= \frac{17}{36}$.

[One may also calculate the probability of the student failing is both the subjects by

$P\left(A_1^C \cap A_2^C\right) = 1 - P\left(A_1 \cup A_2\right) = 1 - \frac{5}{6} = \frac{1}{6}$]

Illustration 1.8 : A publisher has 3 manuscripts in 9 files (with three files for each manuscript) in a closet. If 6 files are chosen at random from the closet, what is the probability that they do not give a whole manuscript?

Solution :

Manuscript M_1	Manuscript M_2	Manuscript M_3
$File_1$ (M_1),	$File_1$ (M_2)	$File_1$ (M_3)
$File_2$ (M_1)	$File_2$ (M_2)	$File_2$ (M_3)
$File_3$ (M_1)	$File_3$ (M_2)	$File_3$ (M_3)

The total number of ways in which 6 files can be drawn at random from the 9 files = ${}^9C_6$

So the sample space is made of ${}^9C_6$ elementary events.

Let A_i (for i = 1, 2, 3) be the event that our of the 6 files selected, 3 files belong to the ith manuscript (i.e. M_i)

Thus, the number of ways in which A_1 is satisfied

= the number of ways in which 3 files of manuscript

M_1 can be picked from the 3 files belonging to

M_1 and remaining 3 files can be picked from the remaining 6 files.

$= {}^3C_3 \times {}^6C_3 = 1 \times \frac{\lfloor 6}{\lfloor 3 \, \lfloor 3} = \frac{4 \times 5 \times 6}{2 \times 3} = 20$

So, $P(A_1) = \frac{20}{{}^{9}C_{6}} = \frac{20}{\frac{\lfloor 9}{\lfloor 6 \lfloor 3}} = \frac{20 \times 2 \times 3}{7 \times 8 \times 9}$

$$= \frac{15}{21}$$

Similarly, $P(A_2) = P(A_3) = \frac{5}{21}$

Further, $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{{}^{3}C_{3} \cdot {}^{3}C_{3}}{{}^{9}C_{6}}$

$$= \frac{1}{\frac{\lfloor 9}{\lfloor 6 \lfloor 3}} = \frac{2 \times 3}{7 \times 8 \times 9}$$

$$= \frac{1}{84}$$

Also, since only 6 files are to be picked, the joint occurence of all the three events A_1, A_2 and A_3 (i.e. 3 files selected from manuscript M_1, 3 files from M_2 and 3 files from M_3) is impossible. So, $P(A_1 \cap A_2 \cap A_3) = 0$

So, the probability that the 6 files chosen at random, gives at least one whole manuscript is given by.

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - \left[P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_2 \cap A_3)\right] + P(A_1 \cap A_2 \cap A_3)$$

$$= \frac{5}{21} + \frac{5}{21} + \frac{5}{21} - \left[\frac{1}{84} + \frac{1}{84} + \frac{1}{84}\right] + 0$$

$$= \frac{15}{21} - \frac{3}{84} = \frac{57}{84}$$

So, the required probability that the 6 files selected do not give a whole manuscript is

$$1 - P\left(A_1 \cup A_2 \cup A_3\right) = 1 - \frac{57}{84} = \frac{27}{84}$$

$$= \frac{9}{28}$$

1.4.2 Theorem of Conditional Probability

Let there be two events A_1 and A_2. A_1 event has non-zero probability. i.e. $P(A_1) > 0$. Let $M(A_1) > 0$. Let $M(A_1)$ be the number of elementary events which are favourable to the event A_1 and $M(A_1 \cap A_2)$ be the number of elementary events that are favourable to both A_1 and A_2. Then, out of the elementary events favourable to A_1, the proportion of the elementary events which are also favourable to A_2 is given by the ratio $\frac{M(\mathbf{A_1 \cap A_2})}{M(A_1)}$

This ratio is known as conditional probability of A_2, given A_1 (i.e. under the given condition that A_1 has already occurred) and is denoted by $P(A_2|A_1)$ or $P_{A_1}(A_2)$

Thus $P(A_1|A_2) = \frac{M(A_1 \cap A_2)}{M(A_1)}$

Similarly, $P(A_1|A_2) = \frac{M(A_1 \cap A_2)}{M(A_2)}$

1.4.3 Theorem of Compound Probability

If A_2 is an event such that $P(A_2) > 0$, (i.e. A_2 has non zero probability) then for any other event A_1, We have $P(A_1 \cap A_2) = P(A_1|A_2) \cdot P(A_2)$

or, $P(A_1|A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)}$

Proof : If the total number of equally likely elementary events in the sample space is M, and $M(A_2)$ and $M(A_1 \cap A_2)$ are the number of elementary events favourable to A_2 and $(A_1 \cap A_2)$ respectively, then using the classical definition of probability we get

$$P(A_1 \cap A_2) = \frac{M(A_1 \cap A_2)}{M}$$

$$= \frac{M(A_1 \cap A_2)}{M(A_2)} \cdot \frac{M(A_2)}{M}$$

$$= P(A_1/A_2) \cdot P(A_2)$$

$\because M(A_2) > 0$

$P(A_1|A_2) = \frac{M(A_1 \cap A_2)}{M(A_2)}$

and

$P(A_2) = \frac{M(A_2)}{M}$

This can be extended for 'n' number of events.
i.e. if there are different events $A_1, A_2 ..., A_n$
such that $P(A_1 \cap A_2 \cap ... \cap A_{n-1}) > 0$, then
$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2|A_1) \times P(A_3|A_1 \cap A_2)$
$P(A_4|A_1 \cap A_2 \cap A_3)...P(A_n|A_1 \cap A_2 \cap ... \cap A_{n-1})$

1.5 Independent Events

Different events are said to be statistically independent, if the probability of occurrence of any of them remains unaffected by the supplementary knowledge regarding the occurrence or non occurrence of any number of the remaining events.

i.e. if there are two events A and B, such that P(A|B) is defined and P(A|B)=P(A), (i.e. the conditional probability of A, given B has already occured is equal to the unconditional probability of A) then it is said that event A is statistically independent of event B.

From the theorem of compound probability, we get $P(A \cap B) = P(A|B).P(B)$.

Now if A and B are statistically independent events, then P(A|B) = P(A)

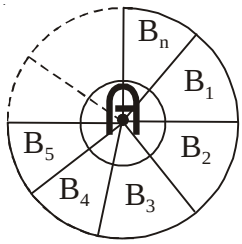
So, $P(A \cap B) = P(A) \cdot P(B)$

Further, if event A is independent of event B, then event B is also independent of event A.

i.e. P(B|A)= P(B)

1.6 Theorem

If there are *n* mutually exclusive and exhaustive events, namely $B_1, B_2...B_n$, such that $P(B_i) > 0$ for each *i*, then $P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A|B_i)$ where event A is an event which can occur only if the set of events $B_1, B_2,...B_n$ occurs.

Proof : 

The inner circle represents the event A.

Due to the fact that the events $B_1, B_2,...B_n$ are mutually exclusive and exhaustive, the events $(A \cap B_1)$, $(A \cap B_2)$, $(A \cup B_3)$,...,$(A \cup B_n)$ are also mutually exclusive events. Further,

$A = (A \cap B_1) \cup (A \cap B_2) \cup ... \cup (A\ B_n)$

So, $P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + ... + P(A \cap B_n)$

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

or, $$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$$ $\begin{cases} \because P(AB_i) = P(A|B_i)P(B_i) \\ \text{for each } i = 1, 2.....n \\ \text{and } P(B_i) > 0 \end{cases}$

1.7 Bayes' Theorem

Let there be '*n*' number of mutually exclusive and exhaustive events, $B_1, B_2..., B_n$ such that none of them has zero probability. Let A be another event which also has non-zero probability. Then for each *i*,

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

Proof : Using the theorem of compound, we can write that

$$P(B_i|A)=\frac{P(A\cap B_i)}{P(A)} \quad [\text{where } P(A) > 0]$$

$$=\frac{P(B_i)P(A|B_i)}{P(A)}$$

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

since, it has been shown that $P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i)$

Illustration 1.9 : Steel pipes are manufactured by three plants A, B, and C with a daily production of 1000, 2000 and 2500 units respectively. It is known from the past experience, that the fractions of the defective pipes produced by the three plants are 0.05, 0.08 and 0.06 respectively. If a pipe, randomly selected from a day's total production is found to be defective, then found out (i) the probability that the defective pipe has been manufactured by plant A. (ii) It has been manufactured by plant B, (iii) It has been manufactured by plant C.

Solution : Let, the event that a pipe is manufactured in plant A be denoted by M_1, the event that a pipe is manufactured in plant Be denoted by M_2, the event that a pipe is manufactured in plant C is denoted by M_3.

The event that a defective pipe is drawn is denoted by D.

Then,

$$P(M_1)=\frac{1000}{1000+2000+2500}=0\cdot1818$$

$$P(M_2)=\frac{2000}{1000+2000+2500}=0\cdot3636$$

$$P(M_3)=\frac{2500}{1000+2000+2500}=0\cdot4545$$

Also, from the given conditions,

$P(D|M_1)=0{\cdot}05\,;\ P(D|M_2)=0{\cdot}08\,;\ P(D|M_3)=0{\cdot}06$

Now, the joint probabilities are given by

$P(D \cap M_1) = P(M_1)\ P(D|M_2)=(0{\cdot}1818)(0{\cdot}05)=0{\cdot}00909$

$P(D \cap M_2) = P(M_2)\ P(D|M_2)=(0{\cdot}3636)(0{\cdot}08)=0{\cdot}029088$

$P(D \cap M_3) = P(M_3)\ P(D|M_3)=(0{\cdot}4545)(0{\cdot}06)=0{\cdot}02727$

$$\therefore P(D) = \sum_{i=1}^{3} P(M_i)\ P(D/M_i) = (0\cdot00909 + 0\cdot029088 + 0\cdot02727) = 0\cdot065448$$

So, the required posterior probabilities are given by

$$P(M_1|D) = \frac{P(M_1)P(D|M_1)}{\sum_{i=1}^{n} P(M_i)P(D|M_i)} = \frac{0\cdot00909}{0.065448} = 0\cdot138889$$ (using Bayes Theorem)

$$P(M_2|D) = \frac{P(M_2)P(D|M_2)}{\sum_{i=1}^{3} P(M_i)P(D|M_i)} = \frac{0{\cdot}029088}{0.065448} = 0{\cdot}44444$$

$$P(M_3|D) = \frac{P(M_3)P(D|M_3)}{\sum_{i=1}^{3} P(M_i)P(D|M_i)} = \frac{0{\cdot}02727}{0.065448} = 0{\cdot}416667$$

Thus, the probability that the defective item has been

(i) manufactured by plant A is 0.138889.

(ii) manufactured by plant B is 0.44444

(iii) manufactured by plant C is 0.416667

The calculations of the posterior probabilities can be shown with the help of the following table :

Table :

Event	Prior Probability $P(M_i)$	Conditional Probability $P(D\|M_i)$	Joint Probability $P(D \cap M_i) = P(M_i)\ P(D\|M_i)$	Posterior Probability $P(M_i\|D) = \frac{P(M_i)\ P(D\|M_i)}{P(D)}$
M_1	0·1818	0·05	$(0.1818)\ (0.05) = P(M_1)\ P(D\|M_1)$	$\frac{0 \cdot 00909}{0 \cdot 065448} = 0 \cdot 138889$
M_2	0·3636	0·08	$(0·3636)\ (0·08) = 0·029088$	$\frac{0 \cdot 029088}{0 \cdot 065448} = 0 \cdot 44444$
M_3	0·4545	0·06	$(0·4545)\ (0·06) = 0·02727$	$\frac{0 \cdot 02727}{0 \cdot 065448} = 0 \cdot 416667$
Total		0·9999 (=1)	$P(D) = 0·065448$	0·9999 (=1)

Illustration 1.10 : four boxes B_1, B_2, B_3 and B_4 contain some black balls and some white balls. The percentage of the total number of balls in these boxes are respectively 30, 15. 45 and 10. The fractions of white balls in the boxes are respectively 0·3, 0·5, 0·1 and 0·2. If a ball is taken out at random and it is found to be white, then what is the probability that it is taken from the box B_3.

Solution : Let B_i denote the event that a ball is taken from the box B_i (i = 1, 2, 3, 4) and let W denote the event that a ball taken out at random is white in colour.

Then from the informations given, we get

$P(B_1) = 0·3$; $P(B_2) = 0·15$; $P(B_3) = 0·45$, $P(B_4) = 0·10$.

Further,

$$P(W|B_1) = 0·3;\ P(W|B_2) = 0.5;\ P(W|B_3) = 0·1;\ P = (W|B_4) = 0·2$$

Then, using the Bayes' theorem, the required probability is

$$P(B_3|W) = \frac{P(B_3)P(W|B_3)}{\sum_{i=1}^{4} P(B_i)\ P(W|B_i)}$$

$$= \frac{(0\cdot45)(0\cdot1)}{(0\cdot3)(0\cdot3)+(0\cdot15)(0\cdot5)+(0\cdot45)(0\cdot1)+(0\cdot10)(0\cdot2)}$$

$$= \frac{0\cdot045}{0\cdot23} = 0\cdot19565$$

Illustration 1.11 : Given that $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{4}$ and $P(A \cap B) = \frac{1}{5}$

(a) Find the values of $P(B^C)$, $P(A \cup B)$, $P(A \cup B)$, $P(A \cup B^C)$ and $P(A^C \cap B^C)$

(b) State whether the events A and B are (i) Equally likely, (ii) independent (iii) mutually exclusive (iv) exhaustive.

Solution : (a) $P(B^C) = 1 - P(B) 1 - \frac{1}{4} = \frac{3}{4}$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{5}}{\frac{1}{4}} = \frac{4}{5}$$

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{3} + \frac{1}{4} - \frac{1}{5} = \frac{23}{60}$

$$P(A \cap B^C) = P(A) - P(A \cap B) = \frac{1}{3} - \frac{1}{5} = \frac{2}{15}$$

So, $P(A \cup B^C) = P(A) + P(B^C) - P(A \cap B^C) = \frac{1}{3} + \frac{3}{4} - \frac{2}{15} = \frac{57}{60}$

and $P(A^C \cap B^C) = 1 - P(A \cup B) = 1 - \frac{23}{60} = \frac{37}{60}$

(b) (i) As $P(A) \neq P(B)$, the events A and B are NOT equally likely.

(ii) $P(A \cap B)=\frac{1}{5}$; But $P(A) \cdot P(B)=\left(\frac{1}{3}\right)\left(\frac{1}{4}\right)=\frac{1}{12}$. Thus $P(A \cap B) \neq P(A) \cdot P(B)$ and hence the events A and B are **not** intdependent.

(iii) $P(A\ B) \neq 0$, So, the events A and B are **NOT** mutually exclusive.

(iv) $P(A\ B) \neq 1$. So the events A and B are Not exhaustive.

1.8 Summary

The term 'Probability' implies chance. The theory of probability deals with the experiments whose outcomes depend on chance and cannot be predicted with full certainty in advance. The fundamental concepts associated with the probability theory are random experiment, event, sample space, mutually exclusive events, exhaustive ecvents, trial, equally likely events, union of events, intersection of events, difference of events etc. As per the classical definition, probability is defined as $P(A) = \frac{\text{Favourable cases to A}}{\text{Exhaustive cases}}$. There are three important theorems of probability. They are—Theorem of total probability, theorem of conditional probability and theorem of compound probability. Different events are said to be statistically independent if the probability of occurrence of any of them remains unaffected by the supplementary knowledge regarding the occurrence or non-occurrence of any number of the remaining events. Bayes' theorem deals with how the probability of an event is affected by new information. It is based on conditional probabilities.

1.9 Self-Assessment Questions

Long Answer Type Question

1. Discuss the theorem of total probability.
2. Explain the theorem of conditional probability.
3. Discuss the theorem of compound probability.

Short Answer Type Questions

1. Explain the classical definition of probability.
2. Write a brief note on Bayes' Theorem.

3. A bag contain 5 white, 6 red and 7 green balls. Three balls are drawn at random. What is te probability that a white, a red and a green ball are drawn.

Objective Type Questions

1. What is a random experiment?
2. What is an event?
3. What do mean by the term ‘sample space’?
4. Define mutually exclusive events.
5. Define exhaustive events.
6. Define equally likely events.
7. What do you understand by ‘Union of events’?
8. What is intersection of events?
9. What is meant by the difference of events?
10. Define independent events.

Unit 2 ❑ Random Variable and Probability Distributions

Structure

2.0 Objectives

After studying the present unit, you will be able to— (i) understand the concept of random variable and (ii) explain the concepts of Binomial distribution, Poisson distribution, Normal distribution, Exponential distribution and their applications.

2.1 Introduction

A finite real valued measurable function defined, on a sample space can be regarded as a random variable and its value is ascertained by the outcome of its experiment. The probability distribution of a random variable along with its corresponding probabilities. The distribution of a discrete random variable and that of a continuous random variable are known as discrete probability distribution and continuous probability distribution respectively. Binomial and Poisson distributions are in the former category while Normal and Exponential distributions are in the latter group. It is expected that in each of these four distributions various values of a random variable are distributed in the population according to some definite probability law.

2.2 Random Variable

In majority of the real life situations, a real number is associated with each elementary event in a sample space. For example, in the throwing of a die, the numbers 1,2,3,...6 are associated with the six possibilities regarding the face that appears on the top. Hence, it may be stated that a function can be defined on the sample space.

Thus a random variable (or a stochastic variable) may be defined as a real valued function defined on the sample space. Corresponding to each value of a random variable (say X) there is a definite probability. A random variable can be either a discrete random variable or a continuous random variable.

(a) Discrete random variable : A random variable which can take a finite number

or a countably infinite number of values. Any finite interval of a discrete random variable contains alsmost a finite number of discrete values.

(b) Continuous random variable : A random variable which can take an uncountably infinite number of values. If X is a random variable and (l, u) is an interval, such that X can take any real value in the interval (l, u), then X is a continuous random variable.

2.2.1 Probability Distribution

The probability distribution of a random variable 'X' is defined as a statement of the possible values of 'x' along with their corresponding probabilities.

Thus, if there is a statement showing the possible values of a discrete random variable X as $x_1, x_2...x_n$, along side their corresponding probability $P_1, P_2,...P_n$, then the statement gives the probability distribution of X.

2.2.2 Probability Mass Function and Discrete Distribution

Let X be a discrete random variable which assume the values $x_1, x_2...x_n$, with probabilities $P_1, P_2,...P_n$ respectively, such that $P(X = x_i) = P_i$, (i = 1, 2, ...n) and $\sum_{i=1}^{n} P_i = 1$.

Values of Variable X→	x_1,	x_2.....x_n	Total
Probabilities →	P_1	P_2.....P_n	$1\left(=\sum_{i=1}^{n} P_i\right)$

Now let there be a real values function $f(x)$, such that

$f(x)$ = Probability that X assumes the value x

= P (X = x)

i.e. $f(x) = \begin{cases} P_i, \text{when } x=x_i \\ O, \text{ Otherwise} \end{cases}$ $\forall$ i = 1, 2,...,n

Here $f(x)$ is called the probability mass function (p.m.f.) of the discrete random variable X. Sometimes it is simply called 'probability function.' This function should satisfy the following conditions :

(a) $f(x_i) \geq 0 \quad \forall\ i = 1,2,3,...n$

(b) $\sum_{i=1}^{n} f(x_i) = 1$

2.2.3 Probability Density Function and Continuous Distribution

Let X be a continuous variable which can take infinite number real values in the interval (l, u). Then its distribution in the infinite population can be represented by dividing its whole range into suitable class intervals and specifying the probability for each interval. For complete representation, a function $f(x)$ is considered, such that for any interval, the probability of X taking values within that interval is given by the integral of $f(x)$ over that interval. Thus the probability of X assuming any value in the infinitely small interval (x, $x + dx$) [here $dx = 0$] is given by

$P(x \leq X \leq x + dx) = f(x)\ dx$

Also, $P\left[l \leq X \leq u\right] = \int_{l}^{u} f(x)dx$

This function $f(x)$ is known as the probability density function (p.d.f) of the continuous random variable X. The function $f(x)$, i.e., p.d.f should satisfy the following conditions :

(a) $f(x) \geq 0$ for any x whatsoever.

(b) $\int_{\alpha}^{\infty} f(x)dx = 1$

2.2.4 Cumulative Distribution Function

The probability that a random variable X assumes a value less or equal to a specified value x is a function of x. This function is called the cumulative distribution function (c.d.f) of the random variable and is generally denoted by $F(x)$.

i.e. $F(x) = P\ (X \leq x)$; and $0 \leq F(x) \leq 1$.

For the continuous probability distribution of X, with p.d.f, $f(x)$ in the interval ($l \leq x \leq u$) the c.d.f is given by $F(x) = \int_{l}^{x} f(\theta)\ d\theta$

$$\begin{bmatrix} \text{This F (K)} = P(x \le K) \\ \sum F(x), \text{when x is continuous} \\ \int_{-\infty}^{k} f(x) \end{bmatrix}$$

One can get the p.d.f $f(x)$ from the c.d.f. $F(x)$ by the operation.

$$f(x) = \frac{d}{dx} F(x)$$

2.2.5 Mathematical Expectation of a Random Variable

(a) Considering discrete random variable : If X is a discrete random variable which can take 'n' values. $x_1, x_2, x_3 ..., x_n$ with probabilities $P_1, P_2, ..., P_n$ respectively, then the mathematical expectation (or mean) of X is defined as

$$E(X) = \mu_x = \sum_{i=1}^{n} x_i P_i = \sum_{i=1}^{n} x_i P[X = x_i]$$

If the p.m.f of the theoretical distribution of a discrete random variable X is $f(x)$, then the expectation of X is

$$E(X) = \mu_x = \sum_{x} x\, f(x)$$

(b) Considering a continuous random variable : Let X be a continuous random variable, and $f(x)$ be its p.d.f. If the interval within which X can assume values, is (l, u),then the expectation of X is given by

$$\mu = E(X) = \int_{l}^{u} x f(x) dx$$

Now if $l \to \infty$ and $u \to \alpha$, then the expectation of X is $\mu = E(X) = \int_{-\alpha}^{\alpha} x\, f(x) dx$

[Note : The expectation of x exists if and only if the integral $\int_{-\alpha}^{\alpha} x f(x) dx$ is absolutely convergent]

Variance of a random variable : Let X be a random variable. Then the variance of X is defined as

Var (X) = E [X – E(X)]^2

Some important truths :

(i) If X = *a*, (a constant) then E(X) = *a* and var (X) = 0.

(ii) If Y = *b*X, then E(Y) = *b*E(X) and var (Y) = b^2 var (X)

(iii) If there are three random variables, X, Y and Z. Such that Z = X + Y, then E(Z) = E(*x*) + E(*y*).

(iv) If Y = *a* + *b*X, then E (Y) = *a* + *b*E (X)

(v) If X and Y are independent random variables, then E(XY) = E(X) E(Y).

Illustration 2.1 : A box contains 2 black and 3 white balls. If two balls are drawn at random, find the mathematical expectation of the number of block balls.

Solution : Let X be the number of black balls obtained among the 2 balls drawn. The possible number of black balls are 0, 1 and 2.

$$\text{Now } P(X=0)=\frac{{}^2C_0 \cdot {}^3C_2}{{}^5C_2}=\frac{1\times\frac{13}{\lfloor 2\ \lfloor 1}}{\frac{\lfloor 15}{\lfloor 2\ \lfloor 3}}=\frac{3}{10}$$

$$P(X=1)=\frac{{}^2C_0 \cdot {}^3C_1}{{}^5C_2}=\frac{2\times 3}{\frac{\lfloor 5}{\lfloor 2 \lfloor 3}}=\frac{6}{10}=\frac{3}{5}$$

$$P(X=2)=\frac{{}^2C_2 \cdot {}^3C_0}{{}^5C_2}=\frac{1\times 1}{\frac{\lfloor 5}{\lfloor 2 \lfloor 3}}=\frac{1}{10}$$

Thus, we get the following table :

x	$P(x)$	$x\,P(x)$
0	$\frac{3}{10}$	0
1	$\frac{3}{10}$	$\frac{3}{10}$
2	$\frac{1}{10}$	$\frac{2}{10}$
Total	1	$\frac{8}{10}=\frac{4}{5}$

So the mathematical expectation of the number of black balls $=E(X)=\sum_{x=0}^{2} xP(X)=\frac{4}{5}$.

Illustration 2.2 : A die is thrown and the number appearing on the uppermost face is observed. If the number appearing on the uppermost face in a throw is 'X', then a man gains Rs 'x' if x is even, and he loses Rs. x if x is odd. Find the expected money earned by the man.

Solution : If X be the gain of the man, then,

$$X = \text{Rs}\begin{cases} x, \text{ when x is even} \\ -x, \text{ if } x \text{ is odd.} \end{cases}$$

Upper most face number	Gains (or losses) X (in Rs.)	$P(x)$	X $P(x)$
1	–1	$\frac{1}{6}$	$-\frac{1}{6}$
2	2	$\frac{1}{6}$	$\frac{2}{6}$

3	–3	$\frac{1}{6}$	$-\frac{3}{6}$
4	4	$\frac{1}{6}$	$\frac{4}{6}$
5	–5	$\frac{1}{6}$	$\frac{-5}{6}$
6	6	$\frac{1}{6}$	$\frac{6}{6}$
Total		1	$\frac{3}{6}=\frac{1}{2}$

Thus, the expected money earned by the man is $\sum XP(x)=\text{Rs.}\frac{1}{2}$

2.2.6 Moments of a Discrete Distribution

If the p.m.f of the theoretical distribution of a discrete random variable X is $f(x)$, then for that distribution.

r-th moment about θ $=\mu'_{r_\theta}=E(X-\theta)^r=\sum_x (x-\theta)^r\, f(x)$

r-th raw moment $=\mu'_r=E(X^x)^r=\sum_x x^r . f(x)$

r-th central moment $=\mu_r=E[X-E(X)]^r=\sum_x [x-E(x)]^r . f(x)$

$$\left[\begin{array}{l}\text{It may be noted that } \mu'_0=\mu_0=1;\ \mu'_1=\mu=E(X);\\ \mu_1=E(X-\mu)=E(X)-E(\mu)=\mu-\frac{\not{n}}{\not{n}}\mu=0\end{array}\right]$$

Further, the central moments can be calculated as :

$$\mu_2 = \mu'_2 - \mu'^2_1$$

$$\mu_3 = \mu'_3 - 3\mu'_2 + 2\mu'^3_1$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$$

2.2.7 Moment Mean and Variance of Continuous Distributions.

If the p.d.f of the continuous probability distribution of a continuous random variable X, assuming values in the interval (l, u) is $f(x)$, then for that distribution

Mean $= \mu = E(X) = \int_l^u x.f(x)dx$

Variance $= \sigma^2 = E\left[X - E(X)\right]^2 = \int_l^u (x-\mu)^2 f(x)dx$

Expectation of any function $\phi(x)$ is given by

$$E\left[\phi(x)\right] = \int_l^u \phi(x).f(x)dx$$

The moments are :

r-th moment about $\theta = \mu'_r = \int_l^u (x-\theta)^r f(x)dx$

r-th raw moment $\mu'_r = \int_l^u x^r.f(x)dx$

r-th central moment = $\mu_r = \int_l^u (x-\mu)^r f(x)dx$

2.3 Binomial Distribution

Let us consider a set of independent trials, such that each Trial can result in either the occurence of an event E, (referred to as a 'success') or the non-occurence of event E (referred to as a 'failure'). For each tiral, the probability of success is the some and

is denoted by '*p*'. Obviously, for each trial, the probability of failure is also the same and denoted by '*q*'. Thus q = 1 – *p*. This kind of a set of trials is known as a set of Bernoulli trials. Now, the probability of getting '*x*' success and hence (*n* – *x*) failures in a set of '*n*' Bernoulli trials in any preassigned order is $p^x q^{n-x}$. However, *x* success (and hence *n* – *x* failures) in *n* any preassigned order is $p^x q^{n-x}$. However, *x* success (and hence *n* – *x* failures) in *n* independent trials may occur in ${}^nC_x$ ways.

Therefore, the probability of getting '*x*' success in '*n*' Bernoulli trials is given by

$$f(x) = {}^nC_x p^x q^{n-x}, \quad \text{for } x = 0, 1, 2, \ldots, n.$$

$$= 0, \text{ Otherwise}$$

Where *n* is a positive interger; $0 < p < 1$, $q = 1 - P$

The function *f*(*x*) is the probability mass function (p.m.f.) of a Binomial distribution (which is a discrete distribution) with parameters '*n*' and '*p*'. It may be noted, that $f(x) \geq 0$ for all *x*.

and $$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} {}^nC_x p^x q^{n-x}$$

$$= (q+p)^n$$

$$= 1$$

2.3.1 Mean of Binomial Distribution

The mean or the first raw moment of binomial distribution is given by

$$\mu_1' = E(X) = \sum_{x=0}^{n} x.f(x)$$

$$= \sum_{x=1}^{n} x\, {}^nC_x p^x q^{n-x} \qquad [\because \text{when } x = 0, \text{ then } x f(x) = 0]$$

$$= \sum_{x=1}^{n} x \cdot \frac{\lfloor n}{\lfloor x \lfloor n-x} p^x q^{n-x}$$

$$= np\sum_{x=1}^{n}\frac{\lfloor n-1}{\lfloor (x-1)-(x-1)}p^{x-1}q^{n-x}$$

$$= np\sum_{x=1}^{n}{}^{(n-1)}C_{(x-1)}p^{x-1}q^{[(n-1)-(x-1)]} \qquad [\text{where } u = x-1$$

$$= np\sum_{u=0}^{n-1}{}^{(n-1)}C_{u}p^{u}q^{[(n-1)-u]}$$

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

Thus, the mean of a binomial distribution (with parameters n) is $\mu = E(X) = np$.

2.3.2 Standard Deviation of Binomial Distribution

$$\text{Var}(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

Now, $E(X^2) = E[X(X-1)+X]$

$$= E[X(X-1)]+E(X)$$

$$= E[X(X-1)]+np \quad \text{.....(i)} \quad \{\because E(X) = np$$

But $E[X(X-1)] = \sum_{x=0}^{n} x(x-1).f(x)$

$$= \sum_{x=0}^{n} x(x-1)^{n}C_{x}p^{x}q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)\frac{\lfloor n q^{n-x}}{\lfloor x \lfloor n-x}p^{x}$$

$$= n(n-1)\sum_{x=2}^{n}\frac{\lfloor n-2}{\lfloor (x-2)-(x-2)}p^{x}q^{n-x}$$

$$= n(n-1)p^2 \sum_{u=0}^{n-2} \frac{\lfloor n-2}{\lfloor u \lfloor (n-2)-u} p^u q^{[n-2)-u]}, \text{ u = x – 2}$$

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

$$= n\ (n-1)\ p^2 \qquad \{\because q+p=1$$

So, from (i)

$E(X^2) = n\ (n-1)\ p^2 + np$

So, var (X) = $E(X^2) - [E(X)^2$

$= n\ (n-1)\ p^2 + np - (np)^2$

$= np\ [np - p + 1 - np]$

$= np\ (1 - p)$

$= npq$

Hence, the standard deviation of X is

$$\sigma = \sqrt{\text{var (X)}} = \sqrt{npq}$$

2.3.3 Some of the Moments of Binomial Distribution

We have already seen that the first moment about zero and the second moment about zero are given by $\mu_1' = E(X) = np$

$$\mu_2' = E(X^2) = n(n-1)p^2 + np$$

The highest moments about zero (the 3rd and the 4th moments) can be shown to be

$$\mu_3' = n(n-1)\ (n-2)p^3 + 3n(n-1)p^2 + np$$

and $\mu_4' = n(n-1)\ (n-2)(n-3)p^4 + \sigma n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$

The central moments are

$$\mu_2 = \mu_2' - \mu_2'^2 = npq$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = npq(q-p)$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 2n^2p^2q^2 + rqp(1-6pq)$$

Now, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{n^2q^2q^2(q-p)^2}{n^3q^3q^3}$

$$= \frac{(q-p)^2}{npq}$$

So the measure for skewness

$$= \gamma_1 = \sqrt{\beta_1} = \frac{(q-p)}{\sqrt{npq}}$$

when $q = p = \frac{1}{2}$, then γ_1 = 0, and the distribution is symmetrical.

When $\frac{1}{2} < p < 1$, then $\gamma_1 > 0$ and the distribution is negatively skew,

The measure for kurtosis is

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

$$= \frac{3n^2p^2q^2 + nqp(1-6pq)}{n^2p^2q^2} - 3$$

$$= 3 + \frac{npq(1-6pq)}{n^2p^2q^2} - 3$$

$$= \frac{(1-6pq)}{npq}$$

From the above measure, it is seen that if $pq = \frac{1}{6}$, then the distribution is

mesokurtic; if $pq > \frac{1}{6}$, then the distribution is platykurtic; and if $pq < \frac{1}{6}$, then the distribution is leptokurtic.

Illustration 2.3 : Five unbiased coins are tossed. Find the probabilities of (i) getting exactly 2 heads, (ii) getting less that 2 heads, (iii) getting at least 2 heads (iv) getting at most 2 heads.

Solution : In this problem

Total number of trials = n = 5

Probability of getting head in each trial = $p = \frac{1}{2}$

Probability of getting tail in each trial = $q = \frac{1}{2}$

(i) Probability of getting exactly 2 heads is given by

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

$$= \frac{\lfloor 5}{\lfloor 2 \, \lfloor 3} \left(\frac{1}{2}\right)^5$$

$$= \frac{10}{32} = \frac{5}{16}$$

$$= 0.3125$$

(ii) The probability of getting less than 2 heads is

$$P(x < 2) = f(0) + f(1) = {}^5C_0 p^0 q^5 + {}^5C_1 p^1 q^4$$

$$= \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 + 5\left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4$$

$$= \frac{1}{2^5} + \frac{5}{2^5}$$

$$= \frac{6}{32} = 0 \cdot 1875$$

(iii) The probability of getting at least 2 heads is

$$P(x \geq 2) = 1 - P(x < 2)$$

$$= 1 - \left[f_{(0)} + f_{(1)} \right]$$

$$= 1 - 0 \cdot 1875 \qquad \begin{cases} \text{from (ii)} \\ f_{(0)} + f_{(1)} = 0 \cdot 1875 \end{cases}$$

$$= 0 \cdot 8215$$

Also, $P(x \geq 2) = P(x = 2) + P(x = 3) + P(x = 4) + P(x = 5)$

$$= f_{(2)} + f_{(3)} + f_{(4)} + f_{(5)}$$

$$= {}^5C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 + {}^5C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 + {}^5C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 + {}^5C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0$$

$$= \left(\frac{1}{2}\right)^5 \left[{}^5C_2 + {}^5C_3 + {}^5C_4 + {}^5C_5 \right]$$

$$= \frac{1}{2^5} \left[\frac{\lfloor 5}{\lfloor 2 \lfloor 3} + \frac{\lfloor 5}{\lfloor 3 \lfloor 3} + \frac{\lfloor 5}{\lfloor 4 \lfloor 1} + 1 \right]$$

$$= \frac{1}{2^5} [10 + 10 + 5 + 1]$$

$$= \frac{26}{32} = 0 \cdot 8215$$

(iv) The probability of getting at most 2 heads is

$$f_{(0)} + f_{(1)} + f_{(2)} = {}^5C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5 + {}^5C_1\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 + {}^5C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3$$

$$= \frac{1}{2^5}\left[{}^5C_0 + {}^5C_1 + {}^5C_2\right]$$

$$= \frac{1}{2^5}[1+5+10]$$

$$= \frac{16}{32} = 0\cdot 5$$

Illustration 2.4 : Suppose that half the population of a town are consumers of rice. 100 investigators are appointed to find out its truth. Each investigator interviews 10 individuals. How many investigators do you expect to repeat that three or less of the people interviewed are consumers of rice.

Solution : In this problem from the given conditions, the probability that an individual of the town is a consumer of rice = $p = \frac{1}{2}$

Total number of trials = the number of interviews taken by each investigator = $n = 10$

So, the probability that three or less of the people interviewed by an investigator are consumers of rice = $f_{(0)} + f_{(1)} + f_{(2)} + f_{(3)}$

$$= {}^{10}C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{10} + {}^{10}C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^9 + {}^{10}C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 + {}^{10}C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7$$

$$= \frac{1}{2^{10}}\left[{}^{10}C_0 + {}^{10}C_1 + {}^{10}C_2 + {}^{10}C_3\right]$$

$$= \frac{1}{2^{10}}[1+10+45+120]$$

$\frac{176}{1024} = 0.171875$.

So, the required expected number of investigators

= 100 (0·171875)

= 17·1875

2.4 Poisson Distribution

Poisson distribution is also a discrete probability distributions. It has the probability mass function (p.m.f)

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{\lfloor x}; \quad \text{for x = 0, 1, 2.....}$$

= 0 Otherwise

Where λ is a positive quantity

$$e = 1 + \frac{1}{\lfloor 1} + \frac{1}{\lfloor 2} + \frac{1}{\lfloor 3} + ... = 2{\cdot}718$$

Note:

$$e^{\lambda} = \frac{\lambda^{0}}{\lfloor 0} + \frac{\lambda^{1}}{\lfloor 1} + \frac{\lambda^{2}}{\lfloor 2} .....$$

$$\sum_{x=0}^{\infty} \frac{\lambda^{x}}{\lfloor x};$$

$$1 = \sum_{x=0}^{\infty} \frac{\lambda^{x} e^{-\lambda}}{\lfloor x}$$

If is seen from the p.m.f that

$f(x) \geq 0$ for all x

and $\sum_{x=0}^{\infty} f(x)=\sum_{x=0}^{\infty}\frac{e^{-\lambda}\lambda^{x}}{\lfloor x}=1$ {Refer the above notes within bracket}

The poisson distribution is a limiting form of the binomial distribution with n $\to \infty$, $p \to 0$, and np = λ (remaining the same)

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} \left[{}^{n}C_{x}p^{x}q^{n-x}\right] = \lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} \left[\frac{n(n-1)(n-2)...(n-x+1)}{\lfloor x}p^{x}(1-p)^{n-x}\right]$$

$$=\frac{1}{\lfloor x}\lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}}\left[1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).....\left(1-\frac{x-1}{n}\right)(np)^{x}\left(1-\frac{np}{n}\right)^{n-x}\right]$$

$$=\frac{\lambda^{x}}{\lfloor x}\lim_{\substack{n\to\infty \\ p\to 0}}\left[1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).....\left(1-\frac{x-1}{n}\right)\right]\frac{\lim\limits_{\substack{n\to\infty \\ p\to 0}}\left(1-\frac{\lambda}{n}\right)^{n}}{\lim\limits_{\substack{n\to\infty \\ p\to 0}}\left(1-\frac{\lambda}{n}\right)^{x}}$$

$$=\frac{\lambda^{x}}{\lfloor x}\frac{e^{-\lambda}}{1}$$

$$\begin{cases}\because \lim\limits_{n\to\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{m-1}{n}\right)=1; \\ \lim\limits_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}=e^{-\lambda} \text{ and } \lim\limits_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{x}=1\end{cases}$$

$$=\frac{e^{-\lambda}\lambda^{x}}{\lfloor x}$$

2.4.1 Mean of Poisson Distribution

The mean or the first raw moment of poisson distribution is given by

$$\mu_1' = E(X) = \sum_{x=0}^{\infty} x.\frac{e^{-\lambda}\lambda^x}{\lfloor x}$$

$$= e^{-\lambda}\sum_{x=1}^{\infty}\frac{x\lambda^x}{\lfloor x}$$

$$= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{\lfloor x-1}$$

$$= \lambda e^{-\lambda}\sum_{u=0}^{\infty}\frac{\lambda^u}{\lfloor u} \qquad \text{[Where } u = x - 1, \text{ as } x \to 1,\ u \to 0\text{]}$$

$$= \lambda e^{-\lambda}\sum_{u=0}^{\infty}\frac{\lambda^u}{\lfloor u}$$

$$= \lambda e^{-\lambda}.e^{\lambda}$$

$$= \lambda$$

2.4.2 Standard Deviation of Poisson Distribution.

Var (X) = σ^2 = E(X^2) – [E(X)]^2.....(i)

Now, E (X^2) = E [X (X –1) + X]

= E[X (X – 1)] + λ...(ii) [$\because$ E(X) = λ]

But, $E[X(X-1) = \sum_{x=0}^{\infty} x(x-1).f(x)$

$$= \sum_{x=2}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{\lfloor x}$$

$$=\lambda^2\sum_{x=2}^{\infty}\frac{e^{-\lambda}\lambda^{x-2}}{\lfloor x-2}$$

$$=\lambda^2\sum_{u=0}^{\infty}\frac{e^{-\lambda}\lambda^{u}}{\lfloor u} \qquad [u=x-2]$$

$$=\lambda^2 e^{-\lambda}\sum_{u=0}^{\infty}\frac{\lambda^{u}}{\lfloor u}$$

$$=\lambda^2 e^{-\lambda}e^{\lambda}$$

$$=\lambda^2$$

Now, from (ii),

$E(X^2)=\lambda^2+\lambda$

So, Var $(X)=E(X^2)-[E(X)]^2$

$$=\lambda^2+\lambda-\lambda^2$$

$$=\lambda$$

Hence, the standard deviation of X is

$\sigma=\sqrt{\text{var}(X)}=\sqrt{\lambda}$

2.4.3 Some of the Moments of Poisson Distribution

The first four moments about zero are

$\mu'_1 = E(X) = \lambda$

$\mu'_2 = E(X^2) = \lambda^2 + \lambda$

$\mu'_3 = \lambda^3 + 3\lambda^2 + \lambda$

$\mu'_4 = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$

The central moments are

$\mu_2 = \mu'_2 - \mu'^2_1 = \lambda$

$$\mu_3 = \mu'_3 - 3\mu'_2 + 2\mu'^3_1 = \lambda$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 3\lambda^2 + \lambda$$

Now, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda}$

So, the measure for skewness

$$= \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$$

As $\lambda > 0$, $\gamma_1 > 0$ and hence, the poisson distribution is pesitively skew.

The measure for kurtosis is

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{3\lambda^2 + \lambda}{\lambda^2} - 3 = 3 + \frac{1}{\lambda} - 3 = \frac{1}{\lambda}$$

As $\lambda > 0$, $\gamma_2 > 0$ and the Poisson distribution is leptokurtic.

Illustration 2.5 : The probability of getting no misprint in a page of a book is 0·12. What is the probability that a page contains more than 3 misprints? (Use poisson distribution).

Solution : Let the number of misprints in a page be denoted by a random variable X. Let us further assume that X follows poisson probability distribution with parameter λ.

i.e. $f(x) = \frac{e^{-\lambda}\lambda^x}{\lfloor x}$

Thus, from the information provided in the questions,

$$f(0) = \frac{e^{-\lambda}\lambda^0}{\lfloor 0} = 0 \cdot 12$$

or, $e^{-\lambda} = 0 \cdot 12$

or, $-\lambda = \log_e 0{\cdot}12 = -2{\cdot}12026354$

or, $\lambda = 2{\cdot}12026354$

$$\therefore f(1) = \frac{e^{-\lambda}\lambda^1}{\lfloor 1} = (0{\cdot}12)(2{\cdot}12026354)$$

$$= 0{\cdot}25443$$

$$f(2) = \frac{e^{-\lambda}\lambda^2}{\lfloor 2} = \frac{(0{\cdot}12)(2{\cdot}12026354)^2}{2}$$

$$= 0.26973$$

$$f(3) = \frac{e^{-\lambda}\lambda^3}{\lfloor 3} = \frac{(0{\cdot}12)(2{\cdot}12026354)^3}{6} = 0{\cdot}190634$$

So, the probability that a page contains more than 3 misprints is

$P(X > 3) = 1 - P(X \leq 3)$

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

$= 1 - [0{\cdot}12 + 0{\cdot}25443 + 0{\cdot}26973 + 0{\cdot}190639]$

$= 1 - 0{\cdot}834794$

$= 0{\cdot}1652$

Illustration 2.6 : A poisson distribution has a double mode at $x = 2$ and $x = 3$. What is the probability that x will take the value 2?

Solution : Let the p.m.f of X be

$$f(x) = \frac{e^{-\lambda}\lambda^2}{\lfloor x}$$

As the distribution has a double mode mode at $x = 2$ and $x = 3$, it is clear that $f(2) = f(3)$

or, $$\frac{e^{-\lambda}\lambda^2}{\lfloor 2} = \frac{e^{-\lambda}\lambda^3}{\lfloor 3}$$

or, $\frac{\lambda^3}{\lambda^2} = \frac{\lfloor 3}{\lfloor 2}$

or, $\lambda = 3$

$\therefore f(2) = \frac{e^{-3}3^2}{\lfloor 2} = \frac{(0 \cdot 049787)}{8} = 0 \cdot 2240$

2.5 Normal Distribution

Normal Distribution (also called Goussion Distribution) is a continuous probability distribution and is regarded as the most important of all the theoretical distributions for continuous variables. If a continuous variable X follows a normal distribution with mean μ and standard deviation σ, then its p.d.f is given by

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

The distribution is denoted by $N(\mu, \sigma^2)$ and its mean is given by—

$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \mu$

The variance of the distribution is given by

Var (X) = $\int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \sigma^2$

From the p.d.f of the normal distribution, it is evident that the distribution is symmetrical about the point x = μ. (i.e. its mean), since

$f(\mu + a) = f(\mu - a) = \frac{1}{\sigma\sqrt{2\pi)}} e^{-\frac{a^2}{2\sigma^2}}$, whatever be the value of a.

As the distribution is symmetrical about μ, its mean is equal to its median. Thus, mean = median = μ.

Further, $f'(\mu) = \frac{df(x)}{dx}|_{x=\mu} = \frac{d}{dx}\left[\frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(-x-\mu)^2}{2\sigma^2}}\right]_{x=\mu}$

$$= \frac{1}{\sqrt{2\pi\sigma}}\left[e^{\frac{(-x-\mu)^2}{2\sigma^2}} \cdot \left\{-\frac{2(x-\mu)}{2\sigma^2}\right\}\right]_{x=\mu} = 0$$

Also, $f''(\mu) = \left.\frac{d^2 f(x)}{dx^2}\right|_{x=\mu} = -\frac{1}{\sqrt{2\pi}\sigma}\frac{d}{dx}\left[e^{-\frac{(x-\mu)^2}{2\sigma^2}}\left\{\frac{x-\mu}{\sigma^2}\right\}\right]_{|x=\mu}$

$$= \frac{-1}{\sqrt{2\pi\sigma}}\left[e^{-\frac{(x-\mu)^2}{2\sigma^2}}\left\{-\left(\frac{x-\mu}{\sigma^2}\right)^2\right\} + e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sigma^2}\right]_{|x=\mu}$$

$$= \frac{-1}{\sqrt{2\pi\sigma}}\left[0 + \frac{1}{\sigma^2}\right] < 0 \ \{\because \sigma > 0, \pi > 0$$

The fact that $f'(\mu) = 0$ and $f''(\mu) < 0$, shows that the mode of the normal distribution is also equal to μ. i.e. $f(x)$ is maximum when $x = \mu$.

Thus, mean = median = mode = μ, for normal distribution.

2.5.1 Some of the Moments of Normal Distribution

All odd order central moments are zero.

i.e. $\mu_{2r+1} = 0$, for all r = 1, 2, 3,...

Hence, $\mu_1 = \mu_3 = 0$.

The even order central moments are given by

$\mu_{2r} = (2r-1)(2r-3)...(5)(3)(1).\sigma^{2r}$.

Hence, the second order and fourth order central moments are given by

$\mu_2 = \sigma^2$, and $\mu_4 = 3\sigma^4$

Now, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{(\sigma^2)^3} = 0$

So, the measure of skewness is

$\gamma_1 = \sqrt{\beta_1} = 0$ (So, the distribution is symmetric)

The measure of kurtosis is

$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{3\sigma^4}{\sigma^4} - 3 = 0$ (So the distribution is mesokurtic)

2.5.2 Standard Normal Distribution

Let $Z = \frac{X - \mu}{\sigma}$, where X is a normal variable with mean μ and standard deviation σ. Then Z will also be a normal variable with mean equal to zero and standard deviation equal to unity. The distribution followed by Z is known as the standard normal distribution and Z is called a standard normal variable or a normal deriate. E(Z) = 0; var(Z) = 1.

The p.d.f of Z is

$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$; $-\infty < z < \infty$

$P[z \le r] = \phi(r) = \int_{-\infty}^{r} \phi(z)\,dz$

Since the distribution is symmetric,

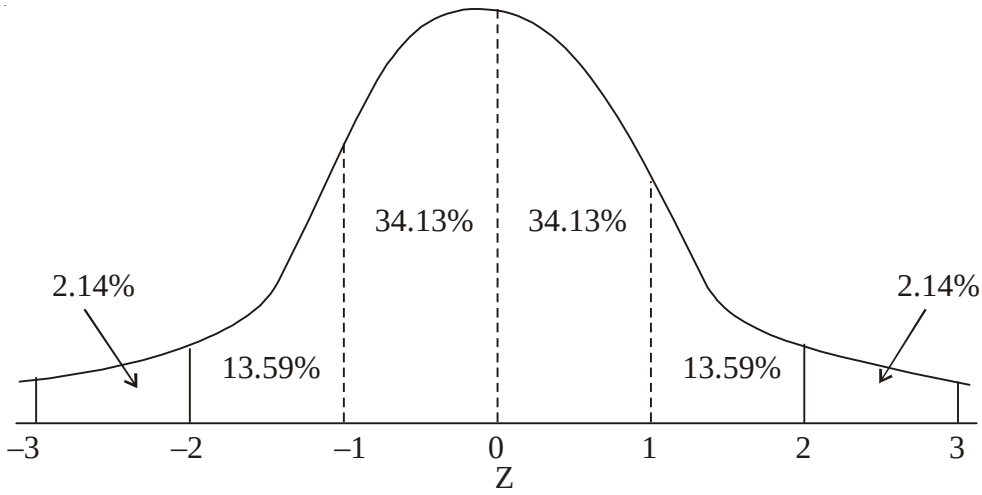
$\phi(z) = \phi(-z)$

$\phi(-r) = 1-\phi(r)$ for any $r > 0$

Also $\phi(-\infty) = 0$; $\phi(0) = 0.5$ and $\phi(\infty) = 1$

$P[r_1 \le z \le r_2] = \phi(r_2) - \phi(r_1) = \int_{r_1}^{r_2} \phi(z)\,dz$

Area under standard normal curve



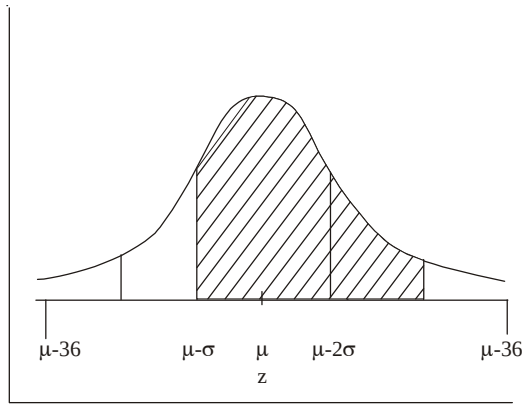
$p[-1 \le z \le 1] = 68.27\%$

$p[-2 \le z \le 2] = 95.45\%$

$p[-3 \le z \le 3] = 99.73\%$

Illustration 2.7 : If a variable X follows normal distribution with mean μ and standard deviation σ, the find out $P[\mu-\sigma<X \le \mu+2\sigma]$

Solution :



$$P[\mu - \sigma < X \le \mu + 2\sigma]$$

$$= P[-1 < \frac{X-\mu}{\sigma} \le +2]$$

$$= P[\frac{X-\mu}{\sigma} \le 2] - P[\frac{X-\mu}{\sigma} \le -1]$$

$$= P[Z \le 2] - P[Z \le -1]$$

$$= \phi(2) - \phi(-1)$$

$$= \phi(2) - [1-\phi(1)]$$

$$= \phi(2) + \phi(1) - 1$$

$$= 0.9772499 + 0.8413447 - 1$$

$$= 0.81859$$

The values of $\phi(2)$ and $\phi(1)$ are taken from standard normal table

Illustration 2.8 : A variable X follows a normal distribution with mean 151 and standard deviation 15. Find the following probabilities:

(i) $P[120 < X \leq 155]$ (ii) $P[X \leq 136]$ (iii) $P[X > 155]$

Solution :

Given : $\mu = 151$ and $\sigma = 15$

(i) $P[120 < X \leq 155]$

$$= P\left[\frac{120-151}{15} < \frac{X-151}{15} \leq \frac{155-151}{15}\right]$$

$$= P[-2.07 < Z \leq 0.27]$$

$$= P[Z \leq 0.27] - P[Z \leq -2.07$$

$$= \phi(0.27) - \phi(-2.07) \qquad \begin{cases}\phi(0.27) = 0.6064 \\ \phi(2.07) = 0.9808\end{cases}$$

$$= \phi(0.27) - [1-\phi(2.07)]]$$

$$= 0.6064 - 1 + 0.9808 = 0.5872$$

(ii) $P[X \leq 136] = P\left[\frac{X-151}{15} \leq \frac{136-151}{15}\right]$

$$= P[Z \leq -1]$$

$$= \phi(-1)$$

$$= 1 - \phi(1) \qquad \{\phi(1) = 0.8413$$

$$= 1 - 0.8413 = 0.1587$$

(iii) $P[X > 155) = 1 - P[X \leq 155]$

$$= 1 - P\left[\frac{X-151}{15} \leq \frac{155-151}{15}\right]$$

$$= 1 - P[Z \leq 0.27]$$

$$= 1 - \phi(0.27) \qquad \{\phi(0.27) = 0.6064$$

$$= 1 - 0.6064 = 0.3936$$

2.6 Exponential Distribution

Exponential distribution is a widely used continuous probability distribution. If a continuous variable X follows exponential distribution with mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$, then its p.d.f is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The parameter $\lambda > 0$. $X \sim$ Exponential (λ)

This distribution is often used to model the time elapsed between events (like service time of customers, inter-arrival time of customers inter-failure time of equipments, etc.) and is found to be the appropriate theoretical distribution for the life time of manufactured products like bulbs, electronic components etc.

The cumulative distribution function (C.D.F) for the exponential distribution is

$$F(x) = \int_0^x f(t)\,dt = \int_0^x \lambda e^{-\lambda t}\,dt$$

$$= [-e^{-\lambda t}]_{t=0}^{t=x}$$

$$= [-e^{-\lambda x} + e^0]$$

$$= 1 - e^{-\lambda x}$$

2.6.1 Mean of Exponential Distribution

The mean of the exponential distribution is given by

$$E(X) = \int_0^\infty x f(x)\,dx = \int_0^\infty x\lambda e^{-\lambda x}\,dx$$

$$= \lambda\left[x\int e^{-\lambda x}\,dx - \int\left\{\frac{d}{dx}\int e^{-\lambda x}\,dx\right\}dx\right]_{x=0}^{x=\infty} \qquad \begin{cases} \because u = u(x) \,\&\, v = v(x) \text{ then} \\ \int uv\,dx = u\int v\,dx - \int\left[\frac{du}{dx}\int v\,dx\right]dx \end{cases}$$

$$= \lambda\left[\frac{-x}{\lambda}e^{-\lambda x} + \int\frac{1}{\lambda}e^{-\lambda x}\,dx\right]_{x=0}^{x=\infty}$$

$$= \frac{\chi}{\chi}\left[-xe^{-\lambda x} - \frac{1}{\lambda}e^{-\lambda x}\right]_{x=0}^{x=\infty}$$

$$= -\left[0 - \frac{1}{\lambda}\right] = \frac{1}{\lambda}$$

2.6.2 Variance of Exponential Distribution

$$E(X^2) = \int_0^{\infty} x^2 f(x)\,dx = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\,dx$$

$$= \lambda\left[x^2 \int e^{-\lambda x}\,dx - \int\left\{\frac{d}{dx}x^2 \int e^{-\lambda x}\,dx\right\}dx\right]_{x=0}^{x=\infty}$$

$$= \lambda\left[\frac{-x^2}{\lambda}e^{-\lambda x} + \int \frac{2x}{\lambda}e^{-\lambda x}\,dx\right]_{x=0}^{x=\infty}$$

$$= \frac{\chi}{\chi}\left[-x^2 e^{-\lambda x} + 2\int xe^{-\lambda x}\,dx\right]_{x=0}^{x=\infty}$$

$$= \left[-x^2 e^{-\lambda x} + 2\left\{\frac{-xe^{-\lambda x}}{\lambda} - \frac{1}{\lambda^2}e^{-\lambda x}\right\}\right]_{x=0}^{x=\infty}$$

$$\left\{\begin{array}{l}\text{Integrating by parts, } \int xe^{-\lambda x}\,dx \\ = \dfrac{-xe^{\lambda x}}{\lambda} - \dfrac{1}{\lambda^2}e^{-\lambda x} + c\end{array}\right.$$

$$= -\left[0 - \frac{2}{\lambda^2}\right] = \frac{2}{\lambda^2}$$

Now, $Var(X) = E(X^2) - [E(X)]^2$

$$= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

So, the standard deviation of $X = \sigma = \sqrt{Var(X)} = \frac{1}{\lambda}$

So, if X ~ Exponential (λ),

then $E(X) = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$

Illustration 2.9: The lifetime (in hours) of an electronic device is a random variable with the following exponential probability density function

$f(x) = 0.025e^{-0.025x}$ for $x \geq 0$

(i) Find the mean life time of the device.

(ii) Find out the probability that the device will fail in the first 20 hours of operation.

(iii) Find out the probability that the device will operate for 80 or more hours before failure.

Solution:

(i) The given distribution of the life time of the electronic device is exponential with parameter $\lambda = 0.025$.

So, the mean life time of the device $= \mu = \frac{1}{\lambda}$

$$= \frac{1}{0.025} = 40 \text{ hours.}$$

(ii) The probability that the device will fail in the first 20 hours of operation

$$= P(X \leq 20)$$

$$= F(20) = \int_0^{20} f(x)\,dx = \int_0^{20} 0.025e^{-0.025x}\,dx$$

$$= [1-e^{-0.025x}]_{x=20}$$

$$= 1-e^{(0.025)(20)}$$

$$= 1-0.606531$$

$$= 0.39347$$

(iii) $P(X > 80) = 1 - P(X \leq 80)$

$$= 1 - F(80)$$

$$= 1 - [1 - e^{-0.025x}]_{x=80}$$

$$= 1 - [1 - e^{(0.025)(80)}]$$

$$= e^{(0.025)(80)}$$

$$= 0.135335$$

Illustration 2.10: The time between arrivals of vehicles at a road crossing follows an exponential distribution with a mean of 15 seconds.

i) Write the p.d.f of the exponential probability distribution followed by the time between arrivals of vehicles.

ii) Find out the probability that the arrival time between vehicles is less or equal to 15 seconds.

iii) Find out the probability that the arricval time between vehicles is less or equal to 8 seconds.

iv) Find out the probability that the arrival time between vehicles is more than 40 seconds.

Solution:

i) In the given problem, the mean time is $E(X) = \mu = 15$ seconds.

So, the parameter $\lambda = \frac{1}{\mu} = \frac{1}{15} = 0.06667$

So, the required p.d.f is

$f(x) = \lambda e^{-\lambda x} = 0.06667 e^{-0.06667x}$; $x \geq 0$

ii) $P[X \leq 15] = F(15)$

$$= \int_0^{15} 0.06667\, e^{-0.06667\, x}\, dx$$

$$= [-e^{-0.06667x}]_{x=0}^{15}$$

$$= 1 - e^{(0.06667)(15)}$$

$$= 1 - 0.36786$$

$$= 0.6321390$$

iii) $P[X \leq 8] = F(8) = 1 - e^{(-0.06667)(8)}$

$$= 1 - 0.586631$$

$$= 0.41337$$

iv) $P[X > 40] = 1 - P[X \leq 40] = 1 - F(40)$

$= 1 - [1 - e^{(-0.06667)(40)}]$

$= e^{(-0.06667)(40)} = 0.06947$

Illustration 2.11: Using the internet provided by ABC Internet Service provider, the average time to download a web page is approximately 25 seconds. Assuming that the time to download a web page follows an exponential distribution, find the following: (i) The probablity that it will take less than 12 seconds to download a web page. (ii) The probability that it will take between 20 seconds and 45 seconds to download a web page.

Solution: It is given that the mean of the concerned exponential distribution is $E(X) = \mu = 25$ seconds.

So, the parameter $\lambda = \frac{1}{\mu} = \frac{1}{25} = 0.04$

So, p.d.f of the exponential distribution is $f(x) = 0.04e^{-0.04x}$; $x \geq 0$

i) $P[X < 12] = F(12)$

$= 1 - e^{(-0.04)(12)}$

$= 1 - 0.6187834$

$= 0.38122$

ii) $P[20 \leq X \leq 45] = F(45) - F(20)$

$= [1 - e^{-(0.04)(45)}] - [1 - e^{-(0.04)(20)}]$

$= e^{-(0.04)(20)} - e^{-(0.04)(45)}$

$= 0.44933 - 0.16530$

$= 0.28403$

2.7 Joint Distribution of Two Random Variables (Discrete)

There are situations in which we require to study more than one random variable at the same time. Let us consider a situation, where we have to study the relationship of two variables, say, 'X' and 'Y'. Let us assume that the random variable X assumes 'k' possible values x_1, x_2, x_3...x_k and corresponding to each value of X, the other

random variable Y assumes '*l*' possible values y_1, y_2, y_3...y_l. Then there are k possible pairs of values (x_i, y_j), where i = 1, 2,..., k and j = 1, 2, ..., *l*. Let P_{ij} denote the probability for each pair of values (x_i, y_j). These P_{ij} give the joint probability distribution of X and Y, as represented in the Table-1 below:

Table 1

Joint Distribution of two random variables, X and Y

Y→ / X↓	y_1	y_2	...	...	...	y_l Total	Marginal
x_1	p_{11}	p_{12}	...	...	...	p_{1l}	p_{10}
x_2	p_{21}	p_{22}	...	...	...	p_{2l}	p_{20}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_k	p_{k1}	p_{k2}	...	...	...	p_{kl}	p_{k0}
Marginal Total	p_{01}	p_{02}	...	...	...	p_{0l}	1

Here, $p_{oj} = P[Y = y_j] = \sum_{i=1}^{k} p_{ij} \quad \forall j = 1, 2, \ldots, l$

and $p_{io} = P[X = x_i] = \sum_{j=1}^{\ell} p_{ij} \quad \forall i = 1, 2, \ldots, k$

The term **Marginal Distribution** of a random variable (say X) refers to the probability distribution of that variable, obtained from the joint distribution irrespective of the values assumed by the other variable (say Y)

Thus p_{io} give the Marginal Probability Distribution of the random variable X.

X :	x_1	x_2	x_3	...	x_k	Total
probability :	p_{10}	p_{20}	p_{30}	...	p_{ko}	1

Similarly, p_{oj} give the marginal distribution of the random variable y.

Y :	y_1	y_2	y_3	...	y_l	Total
probability :	p_{01}	p_{02}	p_{03}	...	p_{0l}	1

From the joint distribution of the two random variables X and Y, the **conditional probability** of the event $X = x_i$, given that $Y = y_j$ has occured (i = 1, 2, ..., k and j = 1,2,...,*l*) can be easily calculated by

$$P\left(X = x_i / Y = y_j\right) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{oj}}$$

Also, $$P\left(Y = y_j / X = x_i\right) = \frac{P(Y = y_j, X = x_i)}{P(X = x_i)} = \frac{p_{ij}}{p_{io}}$$

If 'X' and 'Y' are **satistically independent** random variables, then

$P[X = x_i, Y = y_j] = P[X = x_i]\ P[Y = y_j]$, for all i, for all j

or, $p_{ij} = p_{io} \cdot p_{oj}$; for all i, for all j.

Also, in this case, E(XY) = E(X) E(Y)

and Cov (X,Y) = E(XY) – E(X)E(Y) = 0

[where Cov(X,Y) is the covariance of X and Y]

Thus, if 'X' and 'Y' are **statistically associated**, then $p_{ij} \neq p_{io} \cdot p_{oj}$

Illustration 2.12: An urn contains 3 black, 2 white and 2 red balls. 3 balls are taken at random from the urn. Construct the Joint Distribution of 'The number of black balls' and 'The number of white balls'. Find out the covariance between the number of black balls and the number of white balls obtained.

Solution: Let the number of black balls obtained be X and

the number of white balls obtained be Y.

Then, the possible values of X are 0,1,2,3.

and the possible values of Y are 0,1,2.

Now, $p_{ij} = P[X = x_i, Y = y_j]$

$$= \frac{{}^{3}C_{x_i} \cdot {}^{2}C_{y_j} \cdot {}^{2}C_{3-(x_i+y_j)}}{{}^{7}C_{3}} \qquad \begin{cases} {}^{\forall} x_i = 0,1,2,3 \\ \text{and} \\ y_j = 0,1,2 \end{cases}$$

Table 2

Joint Distribution of the number of black balls (X) and the number of white balls (Y)

X→ / Y↓	0	1	2	3	Marginal total
0	0	$\frac{3}{35}$	$\frac{6}{35}$	$\frac{1}{35}$	$\frac{10}{35}$
1	$\frac{2}{35}$	$\frac{12}{35}$	$\frac{6}{35}$	0	$\frac{20}{35}$
2	$\frac{2}{35}$	$\frac{3}{35}$	0	0	$\frac{5}{35}$
Marginal Total	$\frac{4}{35}$	$\frac{18}{35}$	$\frac{12}{35}$	$\frac{1}{35}$	1

$$p_{00} = \frac{{}^3C_0 . {}^2C_0 . {}^2C_3}{{}^7C_3} = \frac{1 \times 1 \times 0}{35} = 0$$

$$p_{10} = \frac{{}^3C_1 . {}^2C_0 . {}^2C_2}{{}^7C_3} = \frac{3 \times 1 \times 0}{35} = \frac{3}{35}$$

Similarly,

$$p_{20} = \frac{3 \times 1 \times 2}{35} = \frac{6}{35}; p_{30} = \frac{1 \times 1 \times 1}{35} = \frac{1}{35};$$

$$p_{01} = \frac{1 \times 2 \times 1}{35} = \frac{2}{35}; p_{11} = \frac{3 \times 2 \times 2}{35} = \frac{12}{35};$$

$$p_{21} = \frac{3 \times 2 \times 1}{35} = \frac{6}{35}; p_{31} = \frac{1 \times 2 \times 0}{35} = 0;$$

$$p_{02} = \frac{1\times1\times2}{35} = \frac{2}{35}; p_{12} = \frac{3\times1\times1}{35} = \frac{3}{35};$$

$$p_{22} = \frac{3\times1\times0}{35} = 0; p_{31} = \frac{1\times2\times0}{35} = 0;$$

From Table 2, We get

$$E(X) = 1\left(\frac{18}{35}\right) + 2\left(\frac{12}{35}\right) + 3\left(\frac{1}{35}\right) = \frac{45}{35} = \frac{9}{7}$$

$$E(Y) = 1\left(\frac{20}{35}\right) + 2\left(\frac{5}{35}\right) = \frac{30}{35} = \frac{6}{7}$$

$$E(XY) = 1\left(\frac{12}{35}\right) + 2\left(\frac{6}{35}\right) + 3(0) + 2\left(\frac{3}{35}\right) + 0 + 0 = \frac{30}{35} = \frac{6}{7}$$

So, Cov(X,Y) = E(XY) – E(X).E(Y)

$$= \frac{6}{7} - \left(\frac{9}{7}\right)\left(\frac{6}{7}\right) = \frac{-12}{49}$$

2.8 Summary

A random variable may be defined as a finite real valued measurable function defined on a sample space. Random variable may be of two types— discrete random variable and continuous random variable. The probability distribution of a random variable is defined as a statement of the possible values of the variable along with their corresponding probabilities. There are different types of probability distributions. They are Binomial, Poisson, Normal and Exponential distribution. The first two distribution are in the category of discrete probability distribution and the remaining two are continuous probability distribution.

2.9 Self-Assessment Questions

Long Answer Type Question

1. What is Binomial distribution? Explain its mean and standard deviation. Narrate some of the moments of Binomial distribution.
2. What do you mean by 'Poisson Distribution'? Explain the mean, standard deviation and some of the moments of Poisson distribution.
3. What is normal distribution? Explain its properties.
4. What is exponential distribution? Discuss its mean and variance.
5. Six unbiased coins are tossed. Find the probabilities of (i) getting exactly 3 heads, (ii) getting less than 3 heads, (iii) getting at least 3 heads and (iv) getting at most 3 heads.
7. The lifetime (in hours) of an electronic bull is a random variable with the following exponential probability density function

$$f(x) = 0.025e^{-0.025x} \text{ for } x \geq 0$$

   (i) Assesstain the mean lifetime of the bulb.

   (ii) Determine the probability that the bulb will fail in the first 25 hours of operation.

   (iii) Calculate the probability that the bulb will operate for 75 or more hours before failure.

8. There are 10000 electric lamps in the streets of the Hooghly Chinsurah Municipality. If these lamps have an average life of 2000 burning hours with a standard deviation of 230 hours, what number of lamps may be expected to burn for (i) more than 2200 hours, (ii) less than 1800 hours and (ii) between 1600 hours and 2500 hours?

Short Answer Type Questions

1. The mean of a binomial distribution is 40 and its standard deviation is 8. Calculate its n, p and q.
2. "The mean and standard deviation of a poisson distribution are 16 and 9 respectively"—State whether the above statement is ture.

3. In a normal distribution 25% of the items are under 100 and 5% are over 150. Find the mean and standard deviation of the distribution.

Objective Type Questions

1. What is a random variable?
2. What is a discrete random variable?
3. What is a continuous random variable?
4. Define Binomial Distribution.
5. What is Poisson Distribution?
6. What is Normal Distribution?
7. What is Exponential Distribution?

Unit 3 ❑ Sampling Theory

Structure

3.0 Objectives

After studying the present unit, you will be able to— (i) understand the concept of sampling (ii) explain various probability sampling and non-probability sampling techniques (iii) discuss different practical methods of selecting random sample (iv) narrate some important sampling distributions and (v) explain the application of Central Limit Theorem.

3.1 Introduction

The prime objective of any statistical enquiry is to provide complete information on the characteristics of the population. Considering each member of the population in a statistical enquiry is called complete enumenation which is time consuming and laborious. But sampling which represents the selection of a part of the aggregate statistical material with a view to obtaining information about the whole saves time and cost. Sampling theory is the basis of statistical inference as it involves the study of relationship between a population and the samples drawn from it. In the present unit, the issues associated with the concept of sampling, types of sampling, sampling and non-sampling errors, different sampling distributions and Central Limit Theorem are discussed.

3.2 Basic Concept of Sampling

In any statistical enquiry our main interest lies in acquiring information and developing knowledge about an aggregate statistical material (comprising of numerical characteristics of a group of individuals or items), which is known as the population in the context of that enquiry. The population may be made of the electronic products manufactured by a manufacturing plant in a month, the covid affected patients in a state in a week, the high school teachers in a district etc. In majority of the situations, it is not practicable to collect and examine data from each individual member of the population due to several reasons like resource constraints, practical inconvenience, huge population size etc. Thus, in these situations one is left with no other option than to resort to sampling, i.e. examining only a few members of the population.

Sampling may be defined as the selection of a part of an aggregate statistical material (population) with the objective of obtaining information about the whole population.

The aggregate statistical material, from which the sampling is done is called the ***population*** and the part of the population selected by sampling is called a ***sample***.

While doing sampling, the following two basic principles are of great importance:

a) **Validity :** The principle of validity implies that the sampling should be done in such a manner that the objective interpretation of the results in terms of probability can be done.

b) **Optimisation:** The principle of optimisation implies that the sampling should be done in such a manner that either for a given level of cost (incurred for sampling), the level of efficiency (inverse of the sampling variance of the estimator) attained should be maximum or for a given level of efficiency, the cost incurred should be minimum.

3.3 Comparison between Sampling and Complete Enumeration

In complete enumeration (or complete census) the whole population is enumerated or surveyed whereas in sampling only a part of the aggregate statistical material (i.e. a properly selected representative sample) is surveyed. Sample survey is considered to have some advantages over complete enumeration. They are the following:

i) Both in terms of money and man-hours, sample survey generally results in a reduction of total cost, in comparison to complete census. Since the sample size is less than population size, the expected total cost for sample survey is lower than that for complete census although the per unit cost may be higher. Under resource constraint and time constraint, sample survey is preferred.

ii) In some cases, highly trained personnel and expensive equipment may be required for data collection and thus complete enumeration may not be practicable in those cases. Further, in terms of information collected, demographic, geographical and other boundaries, sample survey may provide greater coverage and hence greater scope.

iii) Due to the scope for employing better trained personnel and ensuring better supervision, a sample survey is generally capable of giving data of better quality, in comparison to complete census.

iv) In a sample survey if the sample is properly designed, then it gives an idea

about the magnitude of sampling errors involved in it. However, in case of complete census there is no way to obtain any idea about the magnitude of errors (non-sampling) involved in it.

v) In some cases, due to the nature of the population (like infinite population or hypothetical population), it is not possible to do complete enumeration. Further, if by its nature the enumeration is destructive, then we have to resort to a rather small sample.

3.4 Sampling and Non-sampling Errors

In a sample survey, mainly two types of errors may arise. They are i) Sampling errors and ii) Non-sampling errors.

3.4.1 Sampling Errors

As we have already discussed, sampling is selecting a part of the population with the objective of obtaining information about the whole population. Now, since there is always a chance factor associated with sampling, despite employing appropriate process of selection the results obtained from sample may not exactly coincide with the characteristics of the population. So, some error or discrepancy is inevitable, and this discrepancy between the statistic and the parameter is called the sampling error. For example, although the expected value of the sample mean (a statistic which is an unbiased estimator of the parameter, population mean) is the population mean, it is not that every sample will have that same mean. The difference between the mean of a sample and the mean of the population is caused due to sampling error.

The factors that give rise to sampling errors are— a) Sampling bias and b) Chance factor. Sampling bias may arise due to reasons like the use of defective sampling technique or due to substitution of one convenient member of the population. These biases originate from sampling and are absent in complete enumeration. For example, in a study where a random sampling technique would have been appropriate, but one uses purposive or Judgement sampling then sampling bias would arise. Besides, chance causes may also generate discrepancy between sample characteristics and population characteristics.

3.4.2 Non-Sampling Errors

Procedural biases give rise to non-sampling errors like a) Response error, b) Observational error c) Error arising out of non-response and d) Interviewer's Error

Response error arises out of the responses given by the respondents. Responses may be affected by the respondents' pride and as a result one may under-state one's age or overstate one's income or education. Further, factors like personal likings/ disliking, self-interest of the respondents may also create response error.

Observational error generally arises out of eye-estimation and psychological factors of the observer.

Non-response error commonly arises out of unavailability of the respondent or when the respondent refuses or fails to provide the information.

The interviewer's personal beliefs and prejudices may wrongly affect his/her interpretation and this results in Interviewer's error.

3.5 Types of Sampling

Sampling can be broadly classified into two types.They are subjective sampling and objective sampling. Any sampling type which is dependent on the discretion/ personal judgement of the sampler is called subjective sampling. In case of subjective sampling, the sampler is mostly unaware of the extent to which his sample represents the population and thus, the accuracy of the final estimate is also unknown. When the sampling method does not depend on the sampler's own judgement and is based on a sampling rule then it is called objective sampling.

Objective sampling can be broadly classified into probability sampling, non-probability sampling and mixed sampling.

3.5.1 Probability sampling

The type of objective sampling procedure in which each member of the population gets a definite probability of getting included in the sample is called probability sampling. Thus, for each member there is a definite preassigned probability (which may not be necessarily same for all the members) of being selected.

3.5.2 Non-probability Sampling

This is a type of objective sampling in which there is no probability attached to the mode of selection, although there is a fixed sampling rule.

Probability samples are those based on simple random sampling, systematic sampling, stratified sampling, cluster/area sampling whereas non-probability samples

are those based on convenience sampling, judgement sampling and quota sampling techniques. Important sample designs are briefly discussed below:—

i) ***Deliberate sampling:*** In this sampling method, purposive or deliberate selection of particular units of the universe is done for the constitution of a sample. When on the basis of easy access, the population elements are selected for inclusion in the sample, then the sampling method is called convenience sampling. This is a type of non-probability sampling. In order to collect data related to petrol consumption of public vehicles one may select a fixed number of petrol stations and may conduct interviews at these stations. If the researcher's judgement is used for selecting items of the sample from the population, then the sampling method is called judgement sampling. For example, a judgement sample of doctors might be taken to secure reaction to the adaptation of a new medical treatment.

ii) ***Simple random sampling :*** In this probability sampling method the probability of inclusion of each and every item of the population in the sample is equal. Further, in case of finite universe, each one of the possible samples has the same probability of being selected. For example, if a sample of 400 items is to be selected from a population of 16000 items, then one can write the names or numbers of all the 16000 items on pieces of paper and conduct a lottery, or else one can do random sampling using random number tables. When the population is infinite, then the successive selections of items are independent and the selection of each item in a random sample is controlled by the same probability.

iii) ***Systematic sampling:*** In some situations, the appropriate manner in which sampling can be done is by selecting every 20th entry of a list of individuals, every 15th residence of a housing complex and so on. This type of sampling is known as systematic sampling. In this kind of sampling, at the very beginning, in order to pick up the first unit of the sample random numbers are used and this injects an element of randomness in the procedure. Then an appropriate interval of width 'k' (say) is ascertained. Once the 1st unit has been selected by picking some random point in the list, every kth unit of the population is selected as an unit of the sample until the desired number is secured.

iv) ***Stratified sampling:*** The population is divided/stratified into a number of non-overlapping subpopulations (known as strata) and sample items are selected

from each stratum in this type of sampling. There is not much variance among the items within the same stratum (i.e. each stratum is homogeneous). However, there is significant variance between the items of different strata. If the items selected from each stratum is based on simple random sampling the entire procedure, first stratification and then simple random sampling, is known as stratified random sampling. This technique is applied when the population is a heterogenous group.

v) ***Quota sampling:*** ln Quota sampling (an important form of non-probability sampling),the interviewers are given quota to be filled from different strata. The actual selection (on the basis of the given Quota) of items from different strata for sample is left to the interviewer's own judgement. Here, the cost of sampling is generally much lesser than that of stratified sampling, however the emphasis is given on the interviewer's own judgement rather than randomness. Thus, Quota samples generally happen to be judgement samples rather than random samples. Generally, there is a proportionate relationship between the size of quota for each stratum and the size of that stratum in the population.

vi) ***Cluster sampling and area sampling:*** ln Cluster sampling, initially the population is divided into several groups or clusters. Unlike the stratum (which is made of homogeneous units), here each cluster generally retains the heterogeneity. Then the groups or the clusters (rather than individual elements) are selected at random for inclusion in the sample. Suppose a bank wants to collect a sample of its account holders. It has 30000 account holders. The sample size has to be 750 (say). Now the 30000 account holders may be initially grouped into 200 clusters so that each cluster consists of 150 account holders. Five clusters might then be randomly selected (out of the 200 clusters) for the sample. In area sampling at first the total area is divided into a number of smaller non-overlapping areas, then a number of these smaller areas are randomly selected and all units in these small areas are included in the sample. Area sampling is especially helpful where we do not have the list of the population concerned.

vii) ***Multi-stage sampling:*** This sampling technique may be considered as an extension of the idea of cluster sampling. Under multi-stage sampling, in the first stage, the selection of large primary sampling units such as states may be done, then in the next stage districts, then towns and then in the final stage

certain families within towns may be selected. The procedure is referred to as multi-stage random sampling if random sampling is applied at all stages.

viii) Sequential sampling: In this method of sampling, the final size of the sample is not fixed in advance but is determined according to mathematical decisions on the basis of information created as survey progress. In the context of statistical quality control, this sampling design is used under acceptance sampling plan.

3.6 Practical Methods of Selecting Random Sample

Random samples may be selected using methods like i) Lottery method, ii) Method of using random number table. iii) Method of using Roulette wheel

i) Lottery method: Lottery is done by first constructing a miniature population which can be handled conveniently. Then individuals are drawn one by one from that miniature population which is thoroughly shuffled before every draw. For each sampling unit, a ticket bearing an identification mark (say, serial number) of the sampling unit is prepared and then these tickets are placed in similar containers and thoroughly mixed or randomized.

ii) Random Number Table: Random numbers are the numbers which form some well-known sequence of figures in such a manner that successive figures appear in a perfectly random order, i.e. the successive figures are independent and uncorrelated. Thus, a random number series gives an arrangement (may be linear or rectangular), in which each position is occupied by one of the digits 0,1,2,...,9 and the digit occupying any position is selected at random (out of these ten digits) independently of the digits occupying other positions. Random number table consists of large number of random numbers in sets of four digits arranged in rows and columns. Random number tables are constructed using different sets of random numbers like Tippett's series, Fisher and Yates' series, Kendall and Smith's series etc. Randomness of any series of numbers may be tested using tests like 'Frequency test', 'Serial test', 'Poker test', and 'Gap test'. From any of the random number tables, random numbers may be selected, starting from any row or column of any page. According to the requirement, one can select 1-digited, 2-digited, 3-digited or 4-digited random numbers successively without any break until the required number of random numbers are selected. (see die Example 3 and Example 4)

iii) Method of using Roulette wheel : Roulette wheel is a circular horizontal wheel with numbers (generally from 0 to 36) written on its upper surface along its circumference. It is supposed to be perfectly balanced, clean and fair. A fixed point (a pointer) is chosen on the wheel circumference and the wheel is rotated. The number on the wheel which comes in front of the fixed point when the rotatory wheel stops is chosen. This process is repeated and thus random numbers can be generated. Using these random numbers, random samples can be selected.

3.7 Sampling Distribution, Parameter and Statistic

Let samples of size n be drawn from a population of size N, (n<N). If drawn without replacement, then the possible number of samples that can be drawn is ${}^{N}C_{n}$ and if drawn with replacement, then the possible number of samples that can be drawn is N^{n}.

Now, if we focus on only one character of importance and represent it by variable x, then there will be different measures of the population distribution of x. Any statistical measure based on all units in the population is called a parameter, e.g., population mean, population standard deviation etc. It is a measure that occurs in the population distribution of the variable x.

A corresponding measure for sample, i.e. a statistical measure calculated on the basis of sample observations is called a statistic, e.g., sample mean, sample standard deviation etc.

The population members included in the different possible samples that can be drawn from the population may be different. Thus, the value of a statistic is very likely to vary from one sample to another. These differences/variations in the values of a statistic are called sampling fluctuations. Generally, a statistic varies from sample to sample, but the parameter remains a constant.

The frequency distribution of the statistic that would be obtained if the number of samples (each of the same size, i.e. n) were infinite is called the sampling distribution of the statistic. Rather, sampling distribution of a statistic is the probability distribution of that statistic. Sample statistic like sample mean ($\bar{x}$), sample standard deviation (s), sample proportion of defective (p) etc. will all have their own sampling distributions. A sampling distribution of any statistic may have its mean, standard deviation and moments of higher orders. The standard error of a statisticis the standard deviation

of the sampling distribution of that statistic.

3.7.1 Central Limit Theorem

If the population variance (σ^2) is finite, then the sampling distribution of the sample mean ($\bar{x}$) tends to normality for sufficiently large sample size (n).

3.7.2 Expectation and Standard Error of Sample Mean

Expectation of sample mean is given by $E(\bar{x}) = \mu$ (population mean)

Standard Error (S.E) of the sample mean = The Standard Deviation of the sampling distribution of sample mean. It is given by

$$\text{S.E.} = \begin{cases} \dfrac{\sigma}{\sqrt{n}}, & \text{when the population size is infinite or the random sample is drawn with replacement} \\ \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}, & \text{when the population size is finite \& the random sample is drawn with replacement.} \end{cases}$$

Where, N = Population Size, n = Sample Size, σ = S.D of the Population.

3.7.3 Expectation and Standard Error of Sample Proportion

Let P be the proportion of the units belonging to a certain category in a population. If a random sample of size n be drawn from that population and p represents the proportion of the units belonging to the same category in the sample, then the sampling distribution of p is approximately normal distribution with

Expectation of sample proportion E(p)=P and

Standard Error of sample proportion = The standard deviation of the sampling distribution of p. It is given by

$$\text{S.E.} = \begin{cases} \sqrt{\dfrac{PQ}{n}}, & \text{when the population size is infinite or the random sample is drawn with replacement} \\ \sqrt{\dfrac{PQ}{n}}\sqrt{\dfrac{N-n}{N-1}}, & \text{when the population size is finite \& the random sample is drawn with replacement.} \end{cases}$$

where Q = l–P

3.8 Some Important Distributions

3.8.1 Normal Distribution

Pdf is $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$; $-\infty < x < \infty$ where, σ = S.D of the Population, μ = Mean of the Population

Standard Normal Distribution

Normal distribution with $\mu = 0$ and $\sigma = 1$

Pdf is $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}}$; $-\infty < z < \infty$

3.8.2 χ^2 Distribution

Let $y_1, y_2, ..., y_v$, be mutually independent standard normal variables. Then $\sum_{i=1}^{v} y_i^2$ is called a χ^2 with v degrees of freedom.

Pdf is $f(\chi^2) = \frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} e^{-\frac{\chi^2}{2}} . (\chi^2)^{\frac{v}{2}-1}$; $0 < x^2 < \infty$

3.8.3 t Distribution

If y be a standard normal variable and Y a chi-square (χ^2) with v degrees of freedom, distributed independently of y, then the new variable $\frac{y}{\sqrt{\frac{Y}{v}}}$ is called a t with v degrees of freedom.

Pdf is $f(t) = \frac{1}{v^{\frac{1}{2}}\beta\left(\frac{1}{2}, \frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$, where $-\infty < t < \infty$

3.8.4 F-Distribution

If Y_1 and Y_2 are independently distributed as χ^2 with v_1 and v_2 degrees of freedom, respectively, then the random variable $\frac{Y_1/v_1}{Y_2/v_2}$ is called F with (v_1, v_2) degrees of freedom.

Pdf is $f(F) = \frac{(v_1/v_2)^{v_1/2}}{\beta(v_1/2, v_2/2)} F^{(v_1-2)/2}\left(1+\frac{v_1}{v_2}F\right)^{-(v_1+v_2)/2}$ where $0 < F < \infty$

3.9 Sample Size Decisions

Determining the appropriate size of a sample is a crucial decision in any sampling. Size of the sample should be determined by a researcher keeping in view the following points:

1) **Nature of Population:** Population may be either homogenous for a small sample or heterogenous for a large sample in nature.
2) **Nature of proposed classes:** In case a large number of class groups (groups and sub groups) are to be formed, in order to ensure a reasonable number of items in each class group the sample size should also be large.
3) **The survey concerned:** For technical surveys a small sample and for a general survey a large sample is generally considered to be appropriate.
4) **Type of sampling:** A small random sample is considered to be much superior to a larger but inappropriately selected sample.
5) **Level of accuracy and confidence level:** In order to get higher level of precision the sample size should also be large. Similarly, higher the confidence

level (or lower the level of significance) higher should be the sample size.

6) **Money available:** From the practical point of view, the sample size is obviously dependent on the amount of money available for carrying out the study. For collecting a large sample, sufficient fund should be available.

7) **Other factors:** Size of the population, Population variance, Nature of units, Questionnaire size, Training of investigators, the time available for completion of the study, the conditions under which the sampling is being conducted, are some of the other important factors a researcher should be concerned with while selecting the size of the sample.

A) Sample size while estimating a mean

Let, The population mean = μ

The sample size = n

The standard deviation of the population (as estimated from past experience or trial sample) = σ.

The value of the standard variate at a given level of confidence = z.

The sample mean = $\bar{x}$.

The accectable error = e (= $\bar{x} - \mu$)

Then

$$\frac{\bar{x} - \mu}{\text{Standard error}} = z$$

a) In case, the population size is **infinitely large**, or the sample is drawn **with replacement,** then standard error = $\frac{\sigma}{\sqrt{n}}$.

So, $\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = z$

or, $\frac{e}{\frac{\sigma}{\sqrt{n}}} = z$ or, $\frac{\sigma}{\sqrt{n}} = \frac{e}{z}$ or, $\sqrt{n} = \frac{\sigma z}{e}$ or, $n = \frac{\sigma^2 z^2}{e^2}$

b) In case, the population size is finite (say the population size = N) and the sample is drawn without replacement, then the standard error = $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

So, $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} = z$

or, $\frac{(\bar{x}-\mu)^2}{z^2} = \frac{\sigma^2(N-n)}{n(N-1)}$

or, $n = \frac{\sigma^2}{N-1}\left[\frac{z^2}{(\bar{x}-\mu)^2}\right](N-n)$

or, $n = \frac{\sigma^2 z^2}{(N-1)e^2}N - \frac{\sigma^2 z^2}{(N-1)e^2}n$ {putting $\bar{x}-\mu = e$

or, $n\left[1+\frac{z^2\sigma^2}{e^2(N-1)}\right] = \frac{\sigma^2 z^2}{(N-1)e^2}N$

or, $n\left[\frac{(N-1)e^2+z^2\sigma^2}{e^2(N-1)}\right] = \frac{\sigma^2 z^2 N}{e^2(N-1)}$

or, $n = \frac{z^2\sigma^2 N}{(N-1)e^2+z^2\sigma^2}$

B) Sample size while estimating a proportion

Let, p = Proportion of units belonging to a certain category in the sample.

q = 1 – p = Proportion of units not belonging to the concerned category in the sample.

z = The value of the standard variate at a given level of confidence.

n = The sample size.

e = The acceptable error.

a) In case the population is infinitely large and the proportion in the population is to be estimated, then the standard error of sample proportion = $\sqrt{\frac{pq}{n}}$.

So $\frac{e}{\sqrt{\frac{pq(N-n)}{n(N-1)}}} = z$

or $e^2 = \frac{z^2 pq(N-n)}{n(N-1)}$

or $n(N–1)e^2 = z^2pqN–z^2pqn$

or $n[(N–1)e^2 + z^2pq] = z^2pqN$

or $n = \frac{z^2 pqN}{(N-1)e^2 + z^2 pq}$

Illustration 3.1 : The monthly incomes of 1000 salesmen employed by a company is known to be approximately normally distributed. If the company wants to be 95% confident that the true mean of this months income does not differ by more that 2% of the last month's mean income of Rs. 40000/-. What sample size would be required? Assume that the population standard deviation is Rs. 3500/. Had the population been infinitely large, hour would the appropriate sample size be changed?

Solution: It is given in the problem that

N = 1000

σ = 3500

e = 2% of 40000 = 800

z = 1.96 (as per the table of area under standard normal curve for the given confidence level of 95%).

We know that $n = \frac{z^2\sigma^2 N}{(N-1)e^2 + z^2\sigma^2}$

So here $n = \frac{(1.96)^2(3500)^2(1000)}{(1000-1)(800)^2 + (1.96)^2(3500)^2}$

$= \frac{47059600000}{686419600}$

$= 68.558 \simeq 69.$

Ans. Thus the required sample size is 69.

Had the population been infinitely large, then the appropriate sample size would have been

Illustration 3.2 : What should be the size of the sample if a simple random sample from a population of 5000 items is to be drawn to estimate the percent defective within 2% of the true value with 95% probability? What would be the size of the sample if the population is assumed to be infinite in this problem? [From the past experience, the proportion of defectives in the population is assumed to be p = 0.02]

Solution : In the question it is given that

N = 5000

z = 1.96 (as per table of area under the standard normal curve for the given confidence level of 95%)

e = 0.02

p = 0.02

We know that $n = \frac{z^2pqN}{(N-1)e^2 + z^2pq}$

So here $n = \frac{(1.96)^2(0.02)(1-0.02)(5000)}{(0.02)^2(5000-1)+(1.96)^2(0.02)(1-0.02)}$

$= 181.444 \simeq 181$

So the required sample size is 181.

If the population is assumed to be infinite in this problem, then the appropriate sample size is given by $n = \frac{z^2pq}{e^2} = \frac{(1.96)^2(0.02)(0.98)}{(0.02)^2} = 188.2384 \simeq 188$

Illustration 3.3 : Draw a random sample of size 10 without replacements from a population of 240 firms numbered as F_1, F_2, F_3,..., F_{240}.

Solution : In order to select the random sample, we have to take three-digited numbers from the table of random numbers. Let us take the random numbers row-wise from the begining of the 8th line of the first page. To ensure equal probability for each firm, we shall take the numbers from 001 to 960 (i.e. the largest three-digited number, which is divisible by 240) and ignore the other three-digited numbers. The number is to be divided by 240 and the remainder is to be taken. Obviously, the remainder will vary from 000 to 239. The remainders 001 to 239 will correspond to F_1 to F_{239} respectively and 000 will correspond to F_{240}.

When the sampling is done without replacement, a firm once selected cannot be selected again.

The selected sample is shown below (in Table 1) :

Table 1

Random No. S taken from R Number table	728	709	833	236	325	202	778	001	605	845	010
Remainder on dividing by 240	8	229	113	236	85	202	58	001	125	125 ↓ (Reject)	010
Firms selected	F_8	F_{229}	F_{113}	F_{236}	F_{85}	F_{202}	F_{58}	F_{01}	F_{125}	×	F_{10}

** Had the sampling been done with replacement, in the above example, corresponding to R. Number 845, the remainder 125 would not have been rejected and the firm F_{125} would have been selected Twice. Further F_{10} would not have been selected.

Illustration 3.4 : Draw a random sample of size 8 from a population of 85 units (without replacement).

Solution: The population size is 85, which is a two digited number. So here we would assign a two-digited number to each unit of the population in the following monner.

Unit	1	2	3	4		10	11	12		84	85
Code No.	01	02	03	04		10	11	12		84	85

The code numbers from 86 to 99 and the code number 00 are rejected.

Now when the sampling is done without replacement, the selected sample is

Random Nos taken from R. No. table	72	87	09	83	32	36	32	52	02	77	80	01
Code No.	72	Reject	09	83	32	36	Reject	52	02	77	80	01
Selected Unit of	72	×	09	83	32	36	×	52	02	77	80	01

Here, in this example, had the sampling been done with replacement, the 32nd unit of the population would have been selected twice and the 1st unit of the population (which has been selected at the end) would not have been selected in the sample.

3.10 Summary

The selection of a part of an aggregate statistical material (population) with the objective of obtaining information about the whole population is called sampling. The

part of the population selected by sampling is called a sample. In complete enumeration the whole population is enumerated while in sampling a part of the population is surveyed. The error or discrepancy between the statistic and the parameter which is inevitable is called the sampling error whereas procedural biases result in non-sampling error. Sampling can be classified into two categories—probability sampling and non-probability sampling. Non-probability sampling may be of various types, such as deliberate sampling, simple random sampling, systematic sampling, stratified sampling, quota sampling, cluster sampling, multi-stage sampling, sequential sampling etc. There are three basic methods of selecting random samples. They are Lottery method, Method of using random number table and Method of using roulette wheel. Any statistical measure based on all units in the population is called a parameter whereas a statistical measure based on sample observations is called a statistic. In the Central Limit Theorem, if the population variance is finite, then the sampling distribution of the sample mean tends to normality for sufficiently large sample. Some important sampling distributions are normal distribution Chi-square distribution, t distribution and F distribution.

3.11 Self-Assessment Questions

Long Answer Type Questions

1. Distinguish between—
   (a) Sampling and Complete Enumeration
   (b) Sampling and Non-sampling Errors
2. Discuss the different types of sampling techniques.
3. Narrate the different practical methods of selecting random sample.
4. Discuss the factors to be taken into consideration while determining the size of a sample.

Short Answer Type Questions

1. What is sampling? State its features.
2. Mention the major points of difference between sampling and complete enumeration.
3. State the major points of difference between sampling and non-sampling errors.

4. Write a short note on simple random sampling.
5. State the features of quota sampling.
6. How do you prepare a sample applying stratified sampling?

Objective Type Questions

1. What is sampling?
2. What is sampling error?
3. What is non-sampling error?
4. What is simple random sampling?
5. What do you mean by the term ‘deliberate sampling’?
6. What is systematic sampling?
7. What is multi-stage sampling?
8. What is cluster sampling?
9. What is sequential sampling?
10. What do you understand by the term ‘parameter’?
11. What is statistic?
12. What is Chi-square distribution?

Unit 4 ❑ Multiple Regression Analysis

Structure

4.0 Objectives

After studying the present unit, you will be able to (i) understand the concepts of Partial and Multiple correlations (ii) explain the Multiple linear regression model and (iii) identify the problems associated with the multiple regression and their remedies.

4.1 Introduction

While analyzing the relationship between three or more variables, two issues are addressed. The first one is the assessment of nature and degree of relationship between the variables and the other is the estimation of the value of one variable for specified values of the other variables. The former deals with the analysis of partial and multiple correlations while the latter is concerned with the analysis of multiple regression. This module presents different issues associated with partial correlation, multiple correlation and multiple regression.

4.2 Partial Correlation

Partial correlation coefficient measures the extent of relationship between two

variables in such a way that the effects of other related variables are eliminated. In a trivariate distribution in which each of the variables x_1, x_2 and x_3 has n observations, the partial correlation coefficient between x_1 and x_2 represents the correlation coefficient between $x_1$and x_2 after eliminating the effect of x_3 on x_1 and x_2. It is usually denoted by $r_{12.3}$ and is found by using the following formula:

$$r_{12.3} = \frac{\left(r_{12} - r_{13} \times r_{23}\right)}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

where, r_{12} = Pearson's simple correlation between x_1 and x_2,

r_{13} = Pearson's simple correlation between x_1 and x_3,

r_{23} = Pearson's simple correlation between x_2 and x_3.

Similarly, $r_{13.2}$ is the partial correlation coefficient between $x_1$and x_3. It represents the degree of association between $x_1$and x_3 after eliminating the effect of x_2 on x_1 and x_3. It can be computed by using the following formula:

$$r_{13.2} = \frac{\left(r_{13} - r_{12} \times r_{23}\right)}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{23}^2\right)}}$$

$r_{23.1}$ is the partial correlation coefficient between x_2 and x_3. It represents the degree of relationship between x_2 and x_3 after eliminating the effect of $x_1$on x_2 and x_3. The formula used in computing $r_{23.1}$ is as follows:

$$r_{23.1} = \frac{\left(r_{23} - r_{12} \times r_{13}\right)}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{13}^2\right)}}$$

Partial correlation coefficients lie between +1 and –1 i.e. $-1 \leq (r_{12.3}, r_{13.2}, r_{23.1}) \leq +1$

Illustration 4.1

Using the data given below, computer $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$:

r_{12} = 0.65, r_{13} = 0.6, r_{23} = 0.9. Also interpret the results.

Solution:

$$r_{12.3} = \frac{\left(r_{12} - r_{13} \times r_{23}\right)}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}} = \frac{0.65 - 0.6 \times 0.9}{\sqrt{\left(1 - 0.6^2\right)\left(1 - 0.9^2\right)}}$$

$$= \frac{0.65 - 0.54}{\sqrt{(1 - 0.36)(1 - 0.81)}}$$

$$= \frac{0.11}{\sqrt{0.64 \times 0.19}} = \frac{0.11}{0.3487} = 0.315 \text{ (approx)}$$

It indicates that there is a low degree of positive relationship between x_1 and x_2 after eliminating the influence of x_3 on x_1 and x_2.

$$r_{13.2} = \frac{\left(r_{13} - r_{12} \times r_{23}\right)}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{23}^2\right)}} = \frac{0.6 - 0.65 \times 0.9}{\sqrt{\left(1 - 0.65^2\right)\left(1 - 0.9^2\right)}}$$

$$= \frac{0.6 - 0.585}{\sqrt{(1 - 0.4225)(1 - 0.81)}}$$

$$= \frac{0.15}{\sqrt{0.5775 \times 0.19}} = \frac{0.015}{0.3312} = 0.045 \text{ (approx)}$$

It indicates that there is a very low degree of positive relationship between x_1 and x_3 after eliminating the influence of x_2 on x_1 and x_3.

$$r_{23.1} = \frac{\left(r_{23} - r_{12} \times r_{13}\right)}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{13}^2\right)}} = \frac{0.9 - 0.65 \times 0.6}{\sqrt{\left(1 - 0.65^2\right)\left(1 - 0.6^2\right)}}$$

$$= \frac{0.9 - 0.39}{\sqrt{(1-0.4225)(1-0.36)}}$$

$$= \frac{0.51}{\sqrt{0.5775 \times 0.64}} = \frac{0.51}{0.6079} = 0.839 \text{ (approx)}$$

It reflects that there is a high degree of positive relationship between x_2 and x_3 after eliminating the influence of x_1 on x_2 and x_3.

4.3 Multiple Correlation

When three or more variables are considered simultaneously and we want to study the degree of relationship between one of the variables (considered as dependent variable) on the one hand and the rest two or more variables (considered as independent variables) on the other hand, multiple correlation analysis is made.

In a trivariate distribution in which each of the variables x_1, x_2 and x_3 has n observations, the multiple correlation coefficient of x_1 on x_2 and x_3 is the simple correlation coefficient between x_1 and the joint effect of x_2 and x_3 on x_1. It is usually denoted by $R_{1.23}$ and is found by using the following formula:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2.r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

where, r_{12} = Pearson's simple correlation between x_1 and x_2,

r_{13} = Pearson's simple correlation between x_1 and x_3,

r_{23} = Pearson's simple correlation between x_2 and x_3.

Similarly, $R_{2.13}$ is the multiple correlation coefficient of x_2 on x_1 and x_3. It represents the degree of association between x_2 and the joint influence of x_1 and x_3 on x_2. It can be computed by using the following formula:

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2.r_{12}.r_{13}.r_{23}}{1 - r_{13}^2}}$$

$R_{3.12}$ is the multiple correlation coefficient of x_3 on x_1 and x_2. It shows the extent of relationship between x_3 and the joint effect of x_1 and x_2 on x_3. It can be ascertained by using the following formula:

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2.r_{12}.r_{13}.r_{23}}{1 - r_{12}^2}}$$

The coefficient of multiple correlation varies between +1 and 0 i.e. $0 \leq (R_{1.23}, R_{2.13}, R_{3.12}) \leq +1$

Illustration 4.2

Given $r_{12} = 0.65$, $r_{13} = 0.6$, $r_{23} = 0.9$, compute the values of $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$. Also interpret the value of $R_{3.12}$.

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2.r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{0.65^2 + 0.6^2 - 2 \times 0.65 \times 0.6 \times 0.9}{1 - 0.9^2}}$$

$$= \sqrt{\frac{0.4225 + 0.36 - 0.702}{1 - 0.81}}$$

$$= \sqrt{\frac{0.0805}{0.19}} = 0.651$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2.r_{12}.r_{13}.r_{23}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{0.65^2 + 0.9^2 - 2 \times 0.65 \times 0.6 \times 0.9}{1 - 0.6^2}}$$

$$= \sqrt{\frac{0.4225 + 0.81 - 0.702}{1 - 0.36}}$$

$$= \sqrt{\frac{0.5305}{0.64}} = 0.91$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2.r_{12}.r_{13}.r_{23}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{0.6^2 + 0.9^2 - 2 \times 0.65 \times 0.6 \times 0.9}{1 - 0.65^2}}$$

$$= \sqrt{\frac{0.36 + 0.81 - 0.702}{1 - 0.4225}}$$

$$= \sqrt{\frac{0.468}{0.5775}} = 0.9$$

It indicates that there is a very high degree of relationship between x_3 and the joint influence of x_1 and x_2 on x_3.

4.4 Multiple Regression

If we consider reality, we find that very often the values of one variable are influenced by another single variable, but in most of the cases we see that the values of one variable are influenced by many others, e.g., the dividend paying capability of a company is influenced not only by its profitability but also by its liquidity, availability of profitable investment opportunities, industry trend and so on. In this case, in order to analyse the joint effect of the independent variables on the dependent variable,

multiple regression analysis technique is used. The objective of multiple regression analysis is to estimate the most likely value of the dependent variable from the given values of two or more independent variables.

4.4.1 Multiple Linear Regression Model with Three Variables

In a trivariate distribution, three variables, x_1, x_2 and x_3 are involved. In this case, we have three regression equations which are as follows:

1) Regression equation of x_1 on x_2 and x_3 is given by—

$$\left(x_1 - \overline{x}_1\right) = b_{12.3}\left(x_2 - \overline{x}_2\right) + b_{13.2}\left(x_3 - \overline{x}_3\right)$$

where x_1 is the dependent variable and x_2 and x_3 are the independent variables, $\overline{x}_1$, $\overline{x}_2$ and $\overline{x}_3$ are the arithmetic means of the variables x_1, x_2 and x_3 respectively and $b_{12.3}$ and $b_{13.2}$ are the partial regression coefficients. $b_{12.3}$ shows the change in the value of x_1 for one unit change in the value of x_2 after eliminating the influence of x_3 on x_1 and x_2. Similarly, $b_{13.2}$ indicates the change in the value of x_1 for one unit change in the value of x_3 after eliminating the influence of x_2 on x_1 and x_3. $b_{12.3}$ and $b_{13.2}$ can be found by using the following formulae—

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{\left(r_{12} - r_{13}.r_{23}\right)}{1 - r_{23}^2}$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \times \frac{\left(r_{13} - r_{12}.r_{23}\right)}{1 - r_{23}^2}$$

where, σ_1, σ_2 and σ_3 are the standard deviations of the variables x_1, x_2 and x_3 respectively and r_{12}, r_{13} and r_{23} denote the simple correlation coefficients between x_1 and x_2, between x_1 and x_3 and between x_2 and x_3 respectively.

This equation is used in estimating the most likely value of x_1 for given values of x_2 and x_3.

2) Regression equation of x_2 on x_1 and x_3 is given by—

$$\left(x_2 - \overline{x}_2\right) = b_{21.3}\left(x_1 - \overline{x}_1\right) + b_{23.1}\left(x_3 - \overline{x}_3\right)$$

where x_2 is the dependent variable and x_1 and x_3 are the independent variables, $\overline{x_1}$, $\overline{x_2}$ and $\overline{x_3}$ are the arithmetic means of the variables x_1, x_2 and x_3 respectively and $b_{21.3}$ and $b_{23.1}$ are the partial regression coefficients. $b_{21.3}$ represents the change in the value of x_2 for one unit change in the value of x_1 after eliminating the influence of x_3 on x_1 and x_2. Similarly, $b_{23.1}$ indicates the change in the value of x_2 for one unit change in the value of x_3 after eliminating the influence of x_1 on x_2 and x_3. $b_{21.3}$ and $b_{23.1}$ can be ascertained by using the following formulae—

$$b_{21.3} = \frac{\sigma_2}{\sigma_1} \times \frac{\left(r_{12} - r_{23}.r_{13}\right)}{1 - r_{13}^2}$$

$$b_{23.1} = \frac{\sigma_2}{\sigma_3} \times \frac{\left(r_{23} - r_{12}.r_{13}\right)}{1 - r_{13}^2}$$

This equation is used in estimating the most likely value of x_2 for given values of x_1 and x_3.

3) Regression equation of x_3 on x_1 and x_2 is given by—

$$\left(x_3 - \overline{x_3}\right) = b_{31.2}\left(x_1 - \overline{x_1}\right) + b_{32.1}\left(x_2 - \overline{x_2}\right)$$

where x_3 is the dependent variable and x_1 and x_2 are the independent variables, $\overline{x_1}$, $\overline{x_2}$ and $\overline{x_3}$ are the arithmetic means of the variables x_1, x_2 and x_3 respectively and $b_{31.2}$ and $b_{32.1}$ are the partial regression coefficients. $b_{31.2}$ represents the change in the value of x_3 for one unit change in the value of x_1 after eliminating the influence of x_2 on x_3 and x_1. Similarly, $b_{32.1}$ indicates the change in the value of x_3 for one unit change in the value of x_2 after eliminating the influence of x_1 on x_3 and x_2. $b_{31.2}$ and $b_{32.1}$ can be ascertained by using the following formulae—

$$b_{31.2} = \frac{\sigma_3}{\sigma_1} \times \frac{\left(r_{13} - r_{23}.r_{12}\right)}{1 - r_{12}^2}$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2} \times \frac{\left(r_{23} - r_{13}.r_{12}\right)}{1 - r_{12}^2}$$

This equation is used in estimating the most likely value of x_3 for given values of x_1 and x_2.

Illustration 4.3

In a three-variate multiple correlation analysis, the following results were found—

$\overline{x_1} = 6.8$, $\overline{x_2} = 7$, $\overline{x_3} = 74$

$\sigma_1 = 1$, $\sigma_2 = 0.8$, $\sigma_3 = 9$

$r_{12} = 0.6$, $r_{13} = 0.7$, $r_{23} = 0.65$

(The symbols have their usual significance).

Obtain the multiple regression equations. Also estimate the value of x_3 when $x_1 = 4$ and $x_2 = 8$.

Solution:

In a trivariate distribution, we have three regression equations. Regression equation of x_1 on x_2 and x_3 is—

$$\left(x_1 - \overline{x_1}\right) = b_{12.3}\left(x_2 - \overline{x_2}\right) + b_{13.2}\left(x_3 - \overline{x_3}\right)$$

where $$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{\left(r_{12} - r_{13}.r_{23}\right)}{1 - r_{23}^2}$$ and

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \times \frac{\left(r_{13} - r_{12}.r_{23}\right)}{1 - r_{23}^2}$$

Similarly, regression equation of x_2 on x_1 and x_3 is—

$$\left(x_2 - \overline{x_2}\right) = b_{21.3}\left(x_1 - \overline{x_1}\right) + b_{23.1}\left(x_3 - \overline{x_3}\right)$$

where $b_{21.3} = \frac{\sigma_2}{\sigma_1} \times \frac{\left(r_{12} - r_{23}.r_{13}\right)}{1 - r_{13}^2}$ and

$$b_{23.1} = \frac{\sigma_2}{\sigma_3} \times \frac{\left(r_{23} - r_{12}.r_{13}\right)}{1 - r_{13}^2}$$

Regression equation of x_3 on x_1 and x_2 is—

$$\left(x_3 - \overline{x_3}\right) = b_{31.2}\left(x_1 - \overline{x_1}\right) + b_{32.1}\left(x_2 - \overline{x_2}\right)$$

where $b_{31.2} = \frac{\sigma_3}{\sigma_1} \times \frac{\left(r_{13} - r_{23}.r_{12}\right)}{1 - r_{12}^2}$ and

$$b_{32.1} = \frac{\sigma_3}{\sigma_2} \times \frac{\left(r_{23} - r_{13}.r_{12}\right)}{1 - r_{12}^2}$$

At first we have to compute the values of the partial regression coefficients.

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{\left(r_{12} - r_{13}.r_{23}\right)}{1 - r_{23}^2}$$

$$= \frac{1}{0.8} \times \frac{\left(0.6 - 0.7 \times 0.65\right)}{1 - 0.65^2} = \frac{1}{0.8} \times \frac{\left(0.6 - 0.455\right)}{1 - 0.4225}$$

$$= \frac{1}{0.8} \times \frac{0.145}{0.5775} = 0.31 \text{ (approx)}$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \times \frac{\left(r_{13} - r_{12}.r_{23}\right)}{1 - r_{23}^2}$$

$$= \frac{1}{9} \times \frac{(0.7 - 0.6 \times 0.65)}{1 - 0.65^2} = \frac{1}{9} \times \frac{(0.7 - 0.39)}{0.5775}$$

$$= \frac{1}{9} \times \frac{0.31}{0.5775} = 0.06 \text{ (approx)}$$

$$b_{21.3} = \frac{\sigma_2}{\sigma_1} \times \frac{\left(r_{12} - r_{23}.r_{13}\right)}{1 - r_{13}^2}$$

$$= \frac{0.8}{1} \times \frac{(0.6 - 0.65 \times 0.7)}{1 - 0.7^2} = 0.8 \times \frac{(0.6 - 0.455)}{1 - 0.49}$$

$$= 0.8 \times \frac{0.145}{0.51} = 0.23 \text{ (approx)}$$

$$b_{23.1} = \frac{\sigma_2}{\sigma_3} \times \frac{\left(r_{23} - r_{12}.r_{13}\right)}{1 - r_{13}^2}$$

$$= \frac{0.8}{9} \times \frac{(0.65 - 0.6 \times 0.7)}{1 - 0.7^2} = \frac{0.8}{9} \times \frac{(0.65 - 0.42)}{1 - 0.49}$$

$$= \frac{0.8}{9} \times \frac{0.23}{0.51} = 0.04 \text{ (approx)}$$

$$b_{31.2} = \frac{\sigma_3}{\sigma_1} \times \frac{\left(r_{13} - r_{23}.r_{12}\right)}{1 - r_{12}^2}$$

$$= \frac{9}{1} \times \frac{(0.7 - 0.65 \times 0.6)}{1 - 0.6^2} = 9 \times \frac{(0.7 - 0.39)}{1 - 0.36}$$

$$= 9 \times \frac{0.31}{0.64} = 4.36 \text{ (approx)}$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2} \times \frac{\left(r_{23} - r_{13}.r_{12}\right)}{1 - r_{12}^2}$$

$$= \frac{9}{0.8} \times \frac{(0.65 - 0.7 \times 0.6)}{1 - 0.6^2} = \frac{9}{0.8} \times \frac{(0.65 - 0.42)}{1 - 0.64}$$

$$= \frac{9}{0.8} \times \frac{0.23}{0.64} = 4.04 \text{ (approx)}$$

Now by using the given values of $\overline{x_1}$, $\overline{x_2}$ and $\overline{x_3}$ and the computed values of the partial regression coefficients $b_{12.3}$, $b_{13.2}$, $b_{21.3}$, $b_{23.1}$, $b_{31.2}$ and $b_{32.1}$ the regression equations to be fitted are as follows—

i) Regression equation of x_1 on x_2 and x_3 :

$$\left(x_1 - \overline{x_1}\right) = b_{12.3}\left(x_2 - \overline{x_2}\right) + b_{13.2}\left(x_3 - \overline{x_3}\right)$$

or, $x_1 - 6.8 = 0.31(x_2 - 7) + 0.06\ (x_3 - 74)$

or, $x_1 = 6.8 + 0.31\ x_2 - 2.17 + 0.06\ x_3 - 4.44$

$x_1 = 0.31\ x_2 + 0.06\ x_3 + 0.19$

ii) Regression equation of x_2 on x_1 and x_3 :

$$\left(x_2 - \overline{x_2}\right) = b_{21.3}\left(x_1 - \overline{x_1}\right) + b_{23.1}\left(x_3 - \overline{x_3}\right)$$

or, $x_2 - 7 = 0.23(x_1 - 6.8) + 0.04(x_3 - 74)$

or, $x_2 = 7 + 0.23\ x_1 - 1.564 + 0.04\ x_3 - 2.96$

$x_2 = 0.23x_1 + 0.04x_3 + 2.476$

iii) Regression equation of x_3 on x_1 and x_2

$$\left(x_3 - \overline{x_3}\right) = b_{31.2}\left(x_1 - \overline{x_1}\right) + b_{32.1}\left(x_2 - \overline{x_2}\right)$$

or, $x_3 - 74 = 4.36(x_1 - 6.8) + 4.04\ (x_2 - 7)$

or, $x_3 = 74 + 4.36x_1 - 29.65 + 4.04x_2 - 28.28$

$x_3 = 4.36x_1 + 4.04x_2 + 16.07$

When, $x_1 = 4$ and $x_2 = 8$,

$x_3 = (4.36\times4) + (4.04\times8) + 16.07$

$= 17.44 + 32.32 + 16.07$

$= 65.83$

4.4.2 Problems associated with Multiple Regression and their remedies

There are various problems associated with multiple regression analysis. The major ones are as follows:

(i) Standard error: While estimating regression coefficients the standard deviation of the estimates is considered as standard error. The standard error can be reduced by increasing the size of sample.

(ii) Multicolliearity : If there is a high degree of correlation between the independent variables used in the multiple regression equation, then multicolliearity occurs. As a result, the effect of each of the independent variables cannot be precisely ascertained by using the multiple regression analysis. This problem can be tackled by removing the highly correlated independent variables from the regression equation. It can also be solved by increasing the size of sample because increase in the sample size reduces the extent of relationship between the highly correlated independent variables.

(iii) Autocorrelation : When there is a relationship between the value of a variable in the current period and the values of that variable in the past periods, autocorrelation or serial correlation arises. If autocorrelation is present in the

time series data, then the least squares estimates are considered inefficient in regression equations. For the purpose of reducing autocorrelation necessary steps can be taken. First, if first-order autocorrelation exists, the first difference of the time series data is to be taken and then the values of the parameters are to be estimated. Secondly, the generalized least squares method is to be used to the estimates.

(iv) Heteroscedasticity : If the assumption of constant variance of the random error term in multiple regression model violates, the problem of heteroscedasticity arises. Generally, it occurs when the equation is not properly specified and coefficient estimates vary widely. This problem can be reduced by making different transformations of the variables, such as log transformation, square root transformation, cube root transformation, negative reciprocal transformation etc.

(v) Non-normality of residuals : If the incorrect assumption about residuals distribution is made or the constant variance assumption violates, the problem of non-normality of residuals occurs. This problem can be removed applying the generalized linear model. It can also be reduced by making transformation of the dependent variable which restricts the violation of normality assumption of the residuals.

4.5 Summary

At the time of analyzing the relationship between three or more variables, two issues, such as correlation and regression are addressed. Partial correlation measures the degree of relationship between two variables after eliminating the influence of the other variable(s) on the above mentioned two variables. Multiple correlation studies the degree of relationship between one of the variables (considered as dependent variable) and the joint influence of the rest two or more variables (considered as independent variables). Partial correlation coefficient varies between +1 and –1 whereas multiple correlation coefficient ranges between +1 and 0. Multiple regression analysis technique is used in order to analyse the joint effect of the independent variables on the dependent variable. There are various problems associated with multiple regression analysis. The major ones are the problems of (i) Standard error, (ii) Multicolliearity, (iii) Autocorrelation, (iv) Heteroscedasticity, (v) Non-normality of residuals etc.

4.6 Self-Assessment Questions

Long Answer Type Questions :

1. Define partial correlation and multiple correlation. Distinguish between them.
2. In a study of a random sample of 400 students, the following results are obtained :

$\overline{x}_1 = 78, \quad \overline{x}_2 = 90, \quad \overline{x}_3 = 6$

$\sigma_1 = 10, \quad \sigma_2 = 8, \quad \sigma_3 = 1$

$r_{12} = 0.7, \quad r_{13} = 0.8, \quad r_{23} = 0.93$

Find: a) $r_{12.3}$, b) $r_{13.2}$, c) $r_{23.1}$, d) $R_{1.23}$, e) $R_{2.13}$, f) $R_{3.12}$.

Also interpret the results.

3. Using the same data as given in Question No. 2, you are required to find all the three regression equations. Also estimate the value of x_3 when $x_1 = 88$ and $x_2 = 82$.
4. Given : (i)

$\overline{x}_1 = 120, \quad \overline{x}_2 = 152, \quad \overline{x}_3 = 66$

$\sigma_1 = 5.8, \quad \sigma_2 = 8.2, \quad \sigma_3 = 4.56$

$r_{12} = -0.75, \quad r_{13} = -0.88, \quad r_{23} = 0.91$

Ascertain: (a) $r_{12.3}$, (b) $r_{13.2}$, (c) $r_{23.1}$, (d) $R_{1.23}$, (e) $R_{2.13}$, (f) $R_{3.12}$.

Also interpret the results.

(ii) Estimate the value of x_1 when $x_2 = 148$ and $x_3 = 52$.

5. Examine the validity of the following coefficients :

$r_{12} = -0.46, \quad r_{13} = -0.83, \quad r_{23} = -0.63$

6. Discuss the major problems associated with the multiple regression analysis. Mention the ways by which these problems can be reduced / removed.

Short Answer Type Questions :

1. Distinguish between partial correlation and multiple correlation.
2. Distinguish between partial correlation coefficient and partial regression coefficient.
3. Write a short note on partial correlation coefficient.
4. Write a short note on multiple correlation coefficient.

Objective Type Questions :

1. What is partial correlation?
2. What is multiple correlation?
3. What does partial regression coefficient indicate?
4. If $r_{13.2} = -0.83$, then what does it reflect?
5. If $r_{23.1} = 0.02$, then what does it indicate?
6. If $R_{1.23} = 0.57$, then what does it measure?
7. If $R^2_{2.13} = 0.81$, then what does it imply?
8. If $b_{23.1} = -2.81$, then what does it indicate?

Unit 5 ❑ Theory of Attributes

Structure

5.0 Objectives

Literary, an attribute means a quality or characteristic. The theory of attributes deals with qualitative characteristics which cannot be measured in terms of quantity. The study of attributes, therefore, requires special statistical treatment which is completely different from that of the study of variables. After studying the present unit, you will be able to understand the concept of association of attributes and explain the different methods used in ascertaining the nature of association as well as measuring the degree of the same.

5.1 Introduction

When data are collected on the basis of attribute or attributes, we have statistics commonly termed as statistics of attributes. It is not necessary that the objects may possess only one attribute. In such a situation, your interest may remain in knowing whether the attributes are associated with each other or not. Moreover, you may also be interested in measuring the intensity of association between them.

5.2 Classification of Universe on the Basis of Qualitative Characteristics

In the analysis of statistics relating to attributes the universe (population) is classified on the basis of some qualitative characteristics (phenomena). For example, total number of persons in a town or city (i.e. universe or population) may be classified on the basis of qualitative characteristics as 'skilled' and 'unskilled' or 'intelligent' and 'not intelligent' and so on. The same 'population' ('universe') may be classified with reference to only one qualitative characteristic as stated above or with reference to a combination of two or more qualitative characteristics, e.g. on the basis of 'skill' and

'intelligence' as 'skilled', 'unskilled', 'intelligent', 'not intelligent', 'skilled and intelligent', 'skilled but not intelligent', 'unskilled but intelligent' and 'unskilled as well as not intelligent'.

5.3 Purpose of Classification of Universe

The classification of a universe or population on the basis of two or more qualitative characteristics (attributes) is usually made for the purpose of studying whether the attributes are associated with each other or not. It requires a special type of statistical analysis which is popularly known as 'Analysis of Attributes'.

5.4 Notations Used

For the sake of simplicity and convenience in the analysis of attributes, it is necessary to use certain notations to represent different classes of attributes and their corresponding class frequencies. Usually, Roman capital letters A, B, C, D etc. are used to denote classes representing the presence of attributes (i.e. classes representing 'positive attributes') and Greek letters α (alpha), β (beta), γ (gamma), δ (delta) etc. are used to denote classes representing absence of these attributes. (i.e. classes representing 'negative attributes') respectively. Thus if 'A' represents the attribute 'presence of skill' (the positive attribute), 'α' will denote the attribute 'absence of skill' (the negative attribute). Similarly, we generally use the above mentioned notations to denote two opposite attributes or classes as mentioned below :

Positive Attribute (Class)	**Negative Attribute (Class)**
Presence of skill = A	Absence of skill = α
Presence of intelligence = B	Absence of intelligence = β
Presence of playing ability = C	Absence of playing ability = γ
Presence of criminality = D	Absence of criminality = δ

5.4.1 Manner of Using Notations for Combination of Attributes

The combinations of attributes are denoted by grouping together the letters concerned. For example, 'AB' is the combination of the attributes A and B which represents the simultaneous possession of skill and intelligence. Thus, if only two attributes 'skill' and 'intelligence' are taken into consideration in the study, then by

using the above mentioned notations or symbols their combinations may be denoted in the following manner :

i) Presence of both skill and intelligence = AB

ii) Presence of skill but absence of intelligence = Aβ

iii) Absence of skill but presence of intelligence = αB

iv) Absence of both skill and intelligence = αβ

Similarly ABC, Aβγ, αβC etc. represent the simultaneous possession of three attributes and ABCD, AβCδ etc. denote the simultaneous possession of four attributes.

5.4.2 Manner of Using Notations for Class Frequencies

Different attributes in themselves are called classes and the number of observations assigned to them are known as class frequencies. The class frequencies are denoted by bracketing the class notations. For example, (A) represents the frequencies of A, (AB) represents the number of objects possessing both the attributes A and B, (ABC) indicates the number of objects possessing all the three attributes A, B and C etc.

5.4.3 Types of Class Frequencies

Class frequencies can be classified into three categories :

i) An attribute or a combination of two or more attributes of positive nature is termed as positive class. The frequencies of positive classes are known as positive frequencies. For example, the class frequencies of the type (A), (AB), (ABC) etc. are positive frequencies.

ii) An attribute or a combination of two or more attributes of negative nature is termed as negative class. The frequencies of negative classes are known as negative frequencies. For example, the class frequencies of the type (α), (αβ), (αβγ) etc. are negative frequencies.

iii) The class which contains both positive and negative attributes is termed as contrary class. The frequencies of contrary classes are known as contrary frequencies. For example, the class frequencies of the type (αB), (AβC), (αβC) are called the contrary frequencies.

5.4.4 Number of Classes

The total number of classes can be ascertained by using the following formula :

Total number of classes = 3^n

where n represents the number of attributes studied together.

[Here, the total frequency represented by N is taken as a class.]

Therefore, if only one attribute represented by A is studied, the total number of classes will be 3^1 or 3. The classes are A, α and N. Again if two attributes, A and B are studied together, the total number of classes will be 3^2 or 9. They are N, A, B, α, β, AB, Aβ, αB and $\alpha\beta$. Similarly, if three attributes are studied together, the total number of classes will be 3^3 or 27.

5.4.5 Order of Classes

The term 'order of classes' may be defined as the number of attributes represented at a time in various classes in which the universe is divided. In other words, the order of classes depends upon the number of attributes under study. The universe, without any specification, of attributes, is the class of zero order. A class containing one attribute is called the class of the first order. Similarly, a class containing two attributes is called the class of the second order and so on. Thus, a class represented by n attributes is the class of the nth order.

The order of classes and class frequencies can be presented in the following way:

Order	**Classes**	**Class frequencies**
0	Universe	N
1st	A, B, α, β	(A), (B), (α), (β)
2nd	AB, Aβ, αB, $\alpha\beta$	(AB), (Aβ), (αB), ($\alpha\beta$)

Note: Here, it has been assumed that A and B are the two positive attributes and α and β are the two negative attributes.

5.4.6 Relation between Class Frequencies of Various Orders

All the class frequencies of various orders are not independent of each other and any class frequencies can always be expressed in terms of class frequencies of higher order. If the universe is divided on the basis of an attribute A, there will be two classes A (representing presence of the attribute) and α (representing absence of the attribute)

and the corresponding class frequencies of the first order will be (A)) and (α).

Thus,

$$N = (A) + (\alpha)$$

[Class frequencies of the zero order = class frequencies of the first order].

Again, if one more attribute B is taken into account, each of the class frequencies of the first order as shown above will be divided into two class frequencies of the second order as shown below:

$$(A) = (AB) + (A\beta)$$

$$(\alpha) = (\alpha B) + (\alpha\beta)$$

[Class frequencies of the first order = class frequencies of the second order].

Similarly,

$$(B) = (AB) + (\alpha B)$$

$$(\beta) = (A\beta) + (\alpha\beta)$$

5.4.7 Two by Two Contingency Table or Nine-Square Table

For the purpose of obtaining the nature of the relationship between various class frequencies, in case of two attributes studied together, a table with two rows and two columns is prepared. This is known as 'two by two contingency table'. In addition to the two rows and two columns, one row and one column are included in the table to show the ultimate classes. Thus, this table forms nine squares and accordingly, it is also termed as 'nine-square table'.

Let us now prepare a nine-square table for better understanding.

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	($\alpha\beta$)	(β)
Total	(A)	(α)	N

Here, A = presence of attribute A (positive class),

α = absence of attribute A (negative class),

B = presence of attribute B (positive class) and

β = absence of attribute B (negative class)

Again (AB) = class frequency of 2nd order representing the number of objects possessing both the attribute A and B,

(αB) = class frequency of 2nd order representing the number of objects possessing the attribute B but not the attribute A,

(Aβ) = class frequency of 2nd order representing the number of objects possessing the attribute A but not the attribute B, and

($\alpha\beta$) = class frequency of 2nd order representing the number of objects not possessing both the attributes A and B.

Similarly, (A) = class frequency of 1st order representing the number of objects possessing the attribute A,

(α) = class frequency of 1st order representing the number of objects not possessing the attribute A,

(B) = class frequency of 1st order representing the number of objects possessing the attribute B and

(β) = class frequency of 1st order representing the number of objects not possessing the attribute B.

From the above nine-square table, the following relationships can be obtained:

$$(A) = (AB) + (A\beta)$$

$$(\alpha) = (\alpha B) + (\alpha\beta)$$

$$(B) = (AB) + (\alpha B)$$

$$(\beta) = (A\beta) + (\alpha\beta)$$

$$N = (A) + (\alpha) = (B) + (\beta) = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$$

If some of the class frequencies are given, then you can find out the frequencies of the remaining classes using the relationships mentioned above.

Let us take an illustration for better understanding.

Illustration 5.1

From the following data, find out the missing frequencies:

(AB) = 900, (A) = 2700, N = 9000, (B) = 5400

Solution:

Putting the given values in the nine-square table, we get

	A	α	
B	(AB) = 900	(αB)* = 4500	(B) = 5400
β	(Aβ)* = 1800	(αβ)* = 1800	(β) = 3600
	(A) = 2700	(α)* = 6300	N = 9000

[Here * implies the missing frequencies to be found out]

Now, the frequencies of the classes to be found out are (αB), (Aβ), (αβ), (α) and (β).

i) (B) = (AB) + (αB)

or, (αB) = (B) – (AB) = 5400 – 900 = 4500

ii) (A) = (AB) + (Aβ)

or, (Aβ) = (A) – (AB) = 2700 – 900 = 1800

iii) N = (A) + (α)

or, (α) = N – (A) = 9000 – 2700 = 6300

iv) N = (B) + (β)

or, (β) = N – (B) = 9000 – 5400 = 3600

v) (α) = (αB) + (αβ)

or, (αβ) = (α) – (αB) = 6300 – 4500 = 1800

5.5 Consistency of Data

Consistency of a set of class frequencies may be defined as the property that none of them is negative; otherwise, the data for class frequencies are said to be inconsistent.

Illustration 5.2

From the following two cases find out whether the data are consistent or not—

Case I : (A) = 100, (B) = 150, (AB) = 60, N = 500

Case II : (A) = 220, (B) = 640, (AB) = 260, N = 800

Solution:

Case I : By putting the class frequencies in the nine-square table we get the missing frequencies .

	A	α	
B	(AB) = 60	(αB)* = 90	(B) = 150
β	(Aβ)* = 40	(αβ)* = 310	(β)* = 350
	(A) = 100	(α)* = 400	N = 500

[Here * implies the missing frequencies to be found out]

i) (Aβ) = (A) – (AB) = 100 – 60 = 40

ii) (β) = N – (B) = 500 – 150 = 350

iii) (αβ) = (β) – (Aβ) = 350 – 40 = 310

iv) (αB) = (B) – (AB) = 150 – 60 = 90

v) (α) = N – (A) = 500 – 100 = 400

Since all the class frequencies are positive, we conclude that the given data are consistent.

Case II : By putting the class frequencies in the nine-square table we get the missing frequencies.

	A	α	
B	(AB) = 260	(αB)* = 380	(B) = 640
β	(Aβ)* = - 40	(αβ)* = 200	(β)* = 160
	(A) = 220	(α)* = 580	N = 800

[Here * implies the missing frequencies to be found out]

i) $(A\beta) = (A) - (AB) = 220 - 260 = -40$

ii) $(\alpha) = N - (A) = 800 - 220 = 580$

iii) $(\beta) = N - (B) = 800 - 640 = 160$

iv) $(\alpha B) = (B) - (AB) = 640 - 260 = 380$

v) $(\alpha\beta) = (\alpha) - (\alpha B) = 580 - 380 = 200$

The above table shows that the value of $(A\beta)$ is (–) 40 which is incorrect because no class frequencies can never be negative. Therefore, the given data are inconsistent.

5.6 Independence of Attributes

Two attributes A and B are said to be independent if the presence of one attribute is not influenced by the presence or absence of the other attribute. For example, if criminality and ability to play cricket are independent, the proportion of the cricketers among criminals and non-criminals must be same.

5.7 Association of Attributes

Two attributes A and B are said to be associated only if the presence of one attribute depends on the presence or absence of the other attribute.

5.7.1 Positive Association

Two attributes A and B are said to be positively associated if the presence of one attribute is associated with that of the other and the absence of one attribute is associated with that of the other. Such association is, generally, found between illiteracy and criminality, between educational attainment and employment etc.

5.7.2 Negative Association

Two attributes A and B are said to be negatively associated if the presence of one attribute is associated with the absence of the other. Generally, such association is found between vaccination and attack of small-pox, between literacy and poverty etc.

5.7.3 Complete Association

Two attributes A and B are in complete association if A cannot occur without B, though B may occur without A and vice versa. In other words, A and B are in complete association if either all A's are B's i.e., (AB) = (A) or all B's are A's i.e., (AB) = (B).

5.7.4 Complete Disassociation

Two attributes A and B are in complete disassociation if either no A's are B's i.e., (AB) = 0 or no α's are β's i.e., $(\alpha\beta) = 0$.

5.8 Methods of Studying Association

The following methods are generally, used in order to examine whether two attributes are associated or not :

i) Comparison of observed and expected frequencies method

ii) Proportion method

iii) Yule's coefficient of association

iv) Yule's coefficient of colligation

5.8.1 Comparison of Observed and Expected Frequencies Method

Under this method, the actual number of observations are compared with the expected ones. Expectation is the product of the probability of happening an event and the number of observations. If two attributes A and B are studied in a universe N and the class frequencies of these attributes are (A) and (B), then the expectation of (A) and (B) combined is equal to the product of the joint probability of A and B and the total number of observations N.

Symbolically, the expectation of (AB) = $\frac{(A)}{N} \times \frac{(B)}{N} \times N = \frac{(A)(B)}{N}$

Similarly, the expectation of $(\alpha\beta) = \frac{(\alpha)(\beta)}{N}$,

the expectation of $(A\beta) = \frac{(A)(\beta)}{N}$

the expectation of $(\alpha B) = \frac{(\alpha)(B)}{N}$

A and B are positively associated if $(AB) > \frac{(A)(B)}{N}$. If $(AB) < \frac{(A)(B)}{N}$, then the two attributes A and B are negatively associated. A and B are independent if $(AB) = \frac{(A)(B)}{N}$.

By applying this method the nature of association between any two attributes can only be ascertained.

Illustration 5.3

From the following three cases find out whether the attributes A and B are independent, positively associated or negatively associated:

Case I : (A) = 100, (B) = 150, (AB) = 105, N = 250

Case II : (A) = 120, (AB) = 95, (α) = 150, (αB) = 130

Case III : (AB) = 250, (αB) = 150, (Aβ) = 50, ($\alpha\beta$) = 30

Solution:

Case I : Here $\frac{(A)(B)}{N} = \frac{100 \times 150}{250} = 60$

and (AB) = 105 (given)

Since, (AB) > $\frac{(A)(B)}{N}$, thus A and B are positively associated.

Case II :

	A	α	
B	(AB) = 95	(αB) = 130	(B)
α	(Aβ)	($\alpha\beta$)	(β)
	(A) = 120	(α) = 150	N

N = (A) + (α) = 120 + 150 = 270

(B) = (AB) + (αB) = 95 + 130 = 225

$\frac{(A)(B)}{N} = \frac{120 \times 225}{270} = 100$

Since, (AB) < $\frac{(A)(B)}{N}$, thus A and B are negatively associated.

Case III :

	A	α	
B	(AB) = 250	(αB) = 150	(B)
β	(Aβ) = 50	(αβ) = 30	(β)
	(A)	(α)	N

(A) = (AB) + (Aβ) = 250 + 50 = 300

(B) = (AB) + (αB) = 250 + 150 = 400

N = (AB) + (Aβ) + (αB) + (αβ) = 250 + 50 + 150 + 30 = 480

$$\frac{(A)(B)}{N} = \frac{300 \times 400}{480} = 250$$

Since, $(AB) = \frac{(A)(B)}{N}$, thus A and B are independent.

5.8.2 Proportion Method

Under this method, the proportions of the concerned attributes are compared. If $\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}$ then A and B are positively associated. There is a negative association between A and B when $\frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)}$. If $\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$, then A and B are independent. In this way the nature of association between any two attributes can be determined.

By applying this method the nature of association between any two attributes can only be ascertained.

Illustration 5.4

Find whether A and B are independent, positively associated or negatively associated, in each of the following cases by using proportion method:

Case I : (A) = 47, (B) = 62, (AB) = 32, N = 100

Case II : (A) = 490, (AB) = 294, (α) = 570, (αB) = 380

Case III : (AB) = 192, (αB) = 576, (Aβ) = 36, (αβ) = 108

Solution:

Case I :

	A	α	
B	(AB) = 32	(αB)	(B) = 62
β	(Aβ)	(αβ)	(β)
	(A) = 47	(α)	N = 100

$(A\beta) = (A) - (AB) = 47 - 32 = 15$

$(\beta) = N - (B) = 100 - 62 = 38$

$$\frac{(AB)}{(A)} = \frac{32}{47} = 0.68089 \text{ or } 68.09\%$$

and $$\frac{(A\beta)}{(\beta)} = \frac{15}{38} = 0.3947 \text{ or } 39.47\%$$

Since $\frac{(AB)}{(A)} > \frac{(A\beta)}{(\beta)}$, A and B are positively associated.

Case II :

	A	α	
B	(AB) = 294	(αB) = 380	(B)
β	(Aβ)	(αβ)	(β)
	(A) = 490	(α) = 570	N

$$\frac{(AB)}{(A)} = \frac{294}{490} = 0.6 = 60\%$$

$$\frac{(\alpha B)}{(\alpha)} = \frac{380}{570} = 0.6667 = 66.67\%$$

Since $\frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}$, A and B are negatively associated.

Case III :

	A	α	
B	(AB) = 192	(αB) = 576	(B)
β	(Aβ) = 36	(αβ) = 108	(β)
	(A)	(α)	N

$(A) = (AB) + (A\beta) = 192 + 36 = 228$

$(\alpha) = (\alpha B) + (\alpha\beta) = 576 + 108 = 684$

$$\frac{(AB)}{(A)} = \frac{192}{228} = 0.84211$$

and $$\frac{(\alpha B)}{(\alpha)} = \frac{576}{684} = 0.84211$$

Since $\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$, A and B are independent.

5.8.3 Yule's Coefficient of Association

Yule's coefficient of association is the most popular method of studying association between two attributes. It measures the nature of association between two attributes as well as the degree of association between them. Yule's coefficient of association is usually denoted by the Roman capital letter Q.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

The value of Q varies between +1 and –1 i.e., $1 \geq Q \geq -1$. If the attributes are completely associated with each other (perfect positive association), the value of Q will be +1 and if they are completely dissociated (perfect negative association), the

value of Q will be –1. If the attributes are completely independent of each other, the value of Q will be zero.

Illustration 5.5

Investigate the association between darkness of eye-colour in mother and daughter from the following data using Yule's coefficient of association:

Mothers with dark eyes and daughters with dark eyes: 100

Mothers with dark eyes and daughters with not dark eyes: 158

Mothers with not dark eyes and daughters with dark eyes: 178

Mothers with not dark eyes and daughters with not dark eyes: 1564

Solution:

Let A represent mothers with dark eyes and let B represent daughters with dark eyes. Then α and β will denote mothers with not dark eyes and daughters with not dark eyes respectively.

	A	α	
B	100 (AB)	178 (αB)	100+178 = 278 (B)
α	158 (Aβ)	1564 ($\alpha\beta$)	158+1564 = 1722 (β)
	100+158 = 258 (A)	178+1564 =1742 (α)	278+1722 = 2000 N

Yule's coefficient of association (Q) $= \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$= \dfrac{(100\times1564)-(158\times178)}{(100\times1564)+(158\times178)}$

$= \dfrac{156400-28124}{156400+28124}$

$= \dfrac{128276}{184524} = 0.695$ (approx)

Thus, there is a moderately high degree of positive association between the eye colours of mothers and daughters.

Illustration 5.6

In a population of 500 students, the number of married students is 200. Out of 150 students who failed, 60 belonged to the married group. Using Yule's coefficient of association, find out the extent of association between marriage and failure.

Solution:

Let A represent married students and let B represent students who failed. Then α will denote unmarried students and β will denote students who passed.

	A	α	
B	60 (AB)	150 – 60 = 90 (αB)	150 (B)
β	200 – 60 = 140 (Aβ)	300 – 90 = 210 ($\alpha\beta$)	500 - 150 = 350 (β)
	200 (A)	500 – 200 = 300 (α)	500 N

Yule's coefficient of association (Q) = $\dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$= \dfrac{(60\times 210)-(140\times 90)}{(60\times 210)+(140\times 90)}$

$= \dfrac{12600-12600}{12600+12600} = 0$

Thus, there is no association between A and B. Hence, marriage and failure are independent.

Illustration 5.7

300 students appeared in a test for admission to MBA Programme and 90 of them were successful. 69 students received special coaching and out of them 36 students were successful. Using Yule's coefficient of association estimate the utility of special coaching.

Solution:

Let A represent those who were successful, let B represent those who received special coaching. Thus, α will denote unsuccessful students and β will denote those who did not receive special coaching.

	A	α	
B	(AB) = 36	(αB) = 69 – 36 = 33	(B) = 69
α	(Aβ) = 90 - 36 = 54	($\alpha\beta$) = 231 – 54 = 177	(β) = 300 - 69 = 231
	(A) = 90	(α) = 300 - 90 = 210	N = 300

Yule's coefficient of association (Q) = $\dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$= \dfrac{(36\times177)-(54\times33)}{(36\times177)+(54\times33)}$

$= \dfrac{6372-1782}{6372+1782} = \dfrac{4590}{8154} = 0.56$ (approx)

Thus, there is a moderate degree of positive association between A and B. Hence the special coaching was quite effective for getting success in the admission test.

Illustration 5.8

The following table gives some information relating to three towns of West Bengal:

	Siliguri	Durgapur	Chinsurah
Total number (in '000)	2440	1840	2300
Literates (in '000)	400	470	330
Literate Criminals (in '000)	3	2	2
Illiterate Criminals (in '000)	40	20	24

Compare the degree of association between criminality and illiteracy in each of the three towns.

Solution:

Let A and B represent the number of illiterates and that of criminals respectively. Then α and β will denote the number of literates and that of non-criminals respectively.

Siliguri Durgapur Chinsurah

Siliguri

	A	α	
B	40	3	40+3 = 43
β	2040-40 =2000	400-3 = 397	2440-43 = 2397
	2440-400 = 2040	400	2440

Durgapur

	A	α	
B	20	2	20+2 = 22
β	1370-20 = 1350	1818-1350 = 468	1840-22 =1818
	1840-470 = 1370	470	1840

Chinsurah

	A	α	
B	24	2	24+2 = 26
β	1970-24 = 1946	2274-1946 = 328	2300-26 = 2274
	2300-330 = 1970	330	2300

Yule's coefficient of association (Q) = $\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

Siliguri : $Q_S = \frac{(40\times397)-(2000\times3)}{(40\times397)+(2000\times3)} = \frac{15880-6000}{15800+6000} = \frac{9880}{21880} = 0.45$

Durgapur : $Q_D = \frac{(20\times468)-(1350\times2)}{(20\times468)+(1350\times2)} = \frac{9360-2700}{9360+2700} = \frac{6660}{12060} = 0.55$

Chinsurah : $Q_C = \frac{(24\times328)-(1946\times2)}{(24\times328)+(1946\times2)} = \frac{7872-3892}{7872+3892} = \frac{3980}{11764} = 0.34$

In all the three towns a positive association between criminality and illiteracy is

observed. This positive association is the highest in Durgapur and it is followed by Siliguri and Chinsurah respectively.

5.8.4 Yule's Coefficient of Colligation

It is another measure suggested by Prof. Yule for ascertaining the degree of association between two attributes. It is known as Yule's coefficient of colligation. It has the same properties as possessed by Yule's coefficient of association. Yule's coefficient of colligation is usually denoted by Y.

$$Y = \frac{1-\sqrt{K}}{1+\sqrt{K}}, \text{ where } K = \frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}$$

5.8.5 Relation Between Coefficient of Association and Coefficient of Colligation

The relation between Q and Y is that $Q = \frac{2Y}{1+Y^2}$

Let us now examine the validity of this relation.

$$Y = \frac{1-\sqrt{K}}{1+\sqrt{K}}$$

or, $$Y^2 = \left(\frac{1-\sqrt{K}}{1+\sqrt{K}}\right)^2$$

or, $$1+Y^2 = \frac{\left(1-\sqrt{K}\right)^2}{\left(1+\sqrt{K}\right)^2} + 1$$

or, $$1+Y^2 = \frac{\left(1-\sqrt{K}\right)^2 + \left(1+\sqrt{K}\right)^2}{\left(1+\sqrt{K}\right)^2}$$

or, $$1+Y^2=\frac{1-2\sqrt{K}+K+1+2\sqrt{K}+K}{\left(1+\sqrt{K}\right)^2}$$

or, $$1+Y^2=\frac{2+2K}{\left(1+\sqrt{K}\right)^2}$$

or, $$1+Y^2=\frac{2(1+K)}{\left(1+\sqrt{K}\right)^2}$$

or, $$\frac{2Y}{1+Y^2}=\frac{2\left(\frac{1-\sqrt{K}}{1+\sqrt{K}}\right)}{\frac{2(1+K)}{\left(1+\sqrt{K}\right)^2}}$$ [Putting the values of Y and (1+Y)]

or, $$\frac{2Y}{1+Y^2}=\frac{(1-\sqrt{K})}{\left(1+\sqrt{K}\right)}\times\frac{\left(1+\sqrt{K}\right)^2}{(1+K)}$$

or, $$\frac{2Y}{1+Y^2}=\frac{(1-\sqrt{K})\left(1+\sqrt{K}\right)}{1+K}$$

or, $$\frac{2Y}{1+Y^2}=\frac{1-K}{1+K}$$

or, $$\frac{2Y}{1+Y^2}=\frac{1-\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}{1+\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}$$ [Putting the value of K]

$$\text{or,} \qquad \frac{2Y}{1+Y^2} = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{\frac{(AB)(\alpha\beta)+(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}$$

$$\text{or,} \qquad \frac{2Y}{1+Y^2} = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)} = Q$$

$$Q = \frac{2Y}{1+Y^2} \text{ (Proved)}$$

Illustration 5.9

Using the same data as given in Example 5.7 compute Yule's coefficient of colligation.

Solution:

$$\text{Yule's coefficient of colligation (Y)} = \frac{1-\sqrt{K}}{1+\sqrt{K}}$$

$$\text{where } K = \frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)} = \frac{54\times 33}{36\times 177} = \frac{1782}{6372} = 0.279661$$

$$Y = \frac{1-\sqrt{0.279661}}{1+\sqrt{0.279661}} = \frac{1-0.52882984}{1+0.52882984} = \frac{0.47117016}{1.52882984} = 0.308$$

5.9 Summary

A universe or population is classified on the basis of two or more qualitative characteristics (attributes) for the purpose of studying whether the attributes are associated with each other or not. This leads to analysis of association of attributes. Two attributes are said to be independent if the presence of one attribute does not cause the presence or absence of the other. If the presence of one attribute is associated

with that of the other the association is positive association. On the other hand, if the presence of one attribute is associated with the absence of the other, the two attributes are said to be negatively associated. The nature of association is determined by applying various methods. The more important of these are - comparison of observed and expected frequencies method, proportion method, Yule's coefficient of association, Yule's coefficient of colligation etc. However, the degree of association is measured by any one of the last two. Yule's coefficient of association is the most popular and widely used measure. The value of Yule's coefficient of association (Q) always lies between –1 and +1 i.e. $-1 \leq Q \leq 1$. If Q = +1, the two attributes are said to be completely associated and if Q = –1, they are said to be completely dissociated. The relation between Yule's coefficient of association (Q) and coefficient of colligation (Y) is that $Q = \frac{2Y}{1+Y^2}$

5.10 Self-Assessment Questions

Long Answer Type Questions

1) a) What are the various methods of finding whether two attributes are associated, dissociated or independent?

   b) Distinguish between coefficient of correlation and coefficient of association.

2) Define Yule's coefficient of association (Q) and coefficient of colligation (Y). Establish the following relation between Q and Y :

$$Q = \frac{2Y}{1+Y^2}$$

   What is the range of values for Q?

3) Prepare a nine-square table from the following information:

   N = 250, (α) = 186 , (B) = 60, (AB) = 6.

   Also compute Yule's coefficient of association and interpret the result.

4) The following data relate to the darkness of eye colour in fathers and sons:

   Fathers with dark eyes and sons with dark eyes 150

Fathers with dark eyes and sons with not dark eyes	237
Fathers with not dark eyes and sons with dark eyes	267
Fathers with not dark eyes and sons with not dark eyes	2346

i) State whether the darkness of eye colour in fathers and that in sons are independent, positively associated or negatively associated by using

a) Comparison of observed and expected frequencies method and

b) Proportion method.

ii) Also determine the degree of association between the two attributes by using

a) Yule's coefficient of association and

b) Yule's coefficient of colligation.

5) Arrange the following data in nine-square table and find the unknown class frequencies—

Intelligent fathers with intelligent sons	= 250
Dull fathers with intelligent sons	= 75
Intelligent fathers with dull sons	= 40
Dull fathers with dull sons	= 500

Ascertain whether there is any relationship between intelligent father and that of sons.

6) Find whether A and B are independent, positively associated or negatively associated in the following cases—

i) N = 1000, (A) = 470, (B) = 620, (AB) = 320

ii) (A) = 490, (AB) = 294, (α) = 570, (αB) = 380

iii) (AB) = 256, (αB) = 768, (Aβ) = 48, ($\alpha\beta$) = 144

7) From the data given below, find out if the attributes A and B are independent or positively associated or negatively associated—

i) N = 80, (A) = 60, (B) = 53, (AB) = 35

ii) (A) = 480, (AB) = 294, (α) = 580, (αB) = 380

iii) (AB) = 256, (αB) = 768, (Aβ) = 48, ($\alpha\beta$) = 144

8) Following information are supplied:

	City X	City Y	City Z
No. of literates	40,000	47,000	33,000
Literate criminals	300	200	200
Illiterate criminals	4,000	2,000	2,400
Total no. of persons	244,000	184,000	230,000

Compare the degree of association between criminality and illiteracy in each of the three cities. Can any definite conclusion be drawn from this coefficient?

9) Can vaccination be regarded as a preventive measure of small-pox from the data given below—

i) of 2000 persons in locality exposed to small-pox, 450 in all were attacked,

ii) of 2000 persons 365 had been vaccinated and of these only 50 were attacked.

10) From the data given below, compare the association between literacy and employment in rural and urban area—

	Urban Area	Rural Area
Total no. of adult male	23 lakhs	200 lakhs
Literate male	10 lakhs	40 lakhs
Employed male	5 lakhs	12 lakhs
Literate and employed male	3 lakhs	4 lakhs

11) Find coefficient of association and coefficient of colligation between unemployment and educational attainment for the following result of an urban area:

	Employed	Unemployed
Illiterate or below matric	6,000	400
Matric and above	500	100

12) In an examination at which 500 candidates appeared, boys outnumbered girls by 14% of all candidates. Number of passed candidates exceeded the number

of failed candidates by 300. Boys failing in examination numbered 80. Construct the nine-square table and calculate the coefficient of association between boys and success in the examination.

Short Answer Type Questions:

1) Explain the following terms: i) Order of a class, ii) Positive class frequencies, iii) Negative class frequencies, iv) Relationship between class frequencies of various orders, v) Consistency of given data.

2) Given that (A) = 110, (AB) = 85, (B) = 160 and N = 200; find the other frequencies.

3) Given the following frequencies, find the frequencies of the rest of the classes and the value of N—

   (A) = 300, (αB) = 40, ($\alpha\beta$) = 35, (B) = 200

4) State, with reasons, whether in each of the following cases the data are consistent or not—

   Case I : There is only one attribute A which is being studied in respect of a universe. The class frequency (A) has been shown less than zero.

   Case II : There is only one attributes B which is being studied in respect of a universe. The class frequency (B) has been shown greater than N.

   Case III : There are two attributes A and B which are being studies together in respect of a universe. The class frequency (AB) has been shown less than zero.

5) From the following data, examine whether the data are consistent or not :

   (A) = 600, (B) = 900, (AB) = 360, N = 3000.

6) Examine whether the following data are consistent or not:

   (A) = 70, (B) = 105, (AB) = 98, N = 350.

7) What is meant by association of two attributes? Distinguish between association and correlation as the terms are used in Statistics.

8) When are two attributes said to be: a) positively associated,
   b) negatively associated and
   c) independent?

9) Explain the terms 'complete association' and 'complete dissociation'.

Objective Type Questions

1) What is independence of attributes?
2) What is association of attributes?
3) What is positive association of attributes?
4) What is negative association of attributes?
5) What does the term 'complete association' imply?
6) What do you mean by the term 'complete dissociation'?
7) What is Yule's coefficient of association?
8) What is Yule's coefficient of colligation?

Unit 6 ❑ Test of Hypothesis

Structure

6.0 Objectives

After studying the present unit, you will be able to (i) understand how to develop Null and Alternative hypothesis; (ii) illustrate the knowledge of Type-I and Type-II errors; (iii) explain the critical region, confidence interval at different level of significance (iv)

identify the appropriate test-statistic for testing of hypothesis for different samples (small and large); (v) test hypothesis about population mean when σ is known and unknown.

6.1 Introduction

A test of statistical hypothesis is a statistical procedure which, when the sample values have been obtained, leads to a decision to accept or to reject the hypothesis under consideration. In many cases, we are to make decisions about populations on the basis of sample data. Some information as to the feature of the population or the hypothetical values of the parameters may be available and on the basis of certain rules or criteria we may decide whether the hypothesis is acceptable or not in the light of the sample data collected from the population. This is the problem of hypothesis testing or test of signficance. The theory of hypothesis testing begins with a basic assumption about the parameter of the population. This assumption is termed as hypothesis made on the basis of the sample observations. The validity of a hypothesis will be tested by analysing the sample. The procedure which enables us to decide whether a certain hypothesis is true or not, is called test of hypothesis.

6.1.1 Some Definitions

Statistical hypothesis

A statistical hypothesis is an assertion about the probability distribution of random variables which is verified on the basis of a sample.

Null hypothesis and alternative hypothesis

The null hypothesis is that which is tested for possible rejection under the assumption that it is true. It is denoted by H_0. This hypothesis asserts that there is no difference between population and sample in the matter under consideration.

Any hypothesis which contradicts the null hypothesis is called an alternative hypothesis. It is denoted by H_1.

For example [$H_0 : \mu = \mu_0$ against alternatives a) $H_1 : \mu > \mu_0$

or, b) $H_1 : \mu < \mu_0$

or, c) $H_1 : \mu \neq \mu_0$

Test Statistic

Any statistic is a function of sample observations. Test statistic is a statistic whose computed value determines the final decisions regarding acceptance or rejection of

the null hypothesis. The appropriate test statistic is to be chosen very carefully and knowledge of its sampling distribution under null hypothesis is essential in framing decision rules. If the value of the test statistic falls in the critical region, the null hypothesis is rejected.

Level of significance

This is the probability level, under the null hypothesis, which is employed in defining the critical region. It is generally denoted by the symbol α and is usually taken to be 0.05 or 0.01.

Critical region and acceptance region

The set of values of the test statistic which leads to the rejection of the null hypothesis is known as critical region or rejection region of the test. On .the other hand, the values that lead to the acceptance of the null hypothesis are said to form the acceptance region. Here, we are to test the validity of H_0 against that of H ,at a certain level of significance. In a normal distribution the area under the normal curve outside the ordinates at mean ± 1.96 (s.d.) is only 5%, the probability that the observed value of the statistic differs from the expected value of 1.96 times the standard error or more is 0.05, and the probability of a larger difference will be still smaller.

Let $Z = \dfrac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$ and then if $|Z| \geq 1.96$, we reject the null hypothesis.

Therefore $|Z| \geq 1.96$ constitutes the critical region of the test. It is denoted by w, Thus $|Z| \leq 1.96$ constitutes the acceptance region of the test and it is denoted by w.

Type I and Type II error

Probability of type I error is defined as the probability of rejecting the null hypothesis when it is true. The critical region is so determined that the probability of type I error does not exceed the level of significance of the test.

Probability of Type I error = $P(x \in w \mid H_0)$

Probability of type II error is defined as the probability of accepting the null hypothesis when it is really false.

Probability of type II error = $P(x \in \bar{w} \mid H_1) = 1 - P(x \in w / H_1)$.

Power of a test

The power of a test is defined as the probability of rejecting the null hypothesis when it is false. On the other hand, power of a test is defined as

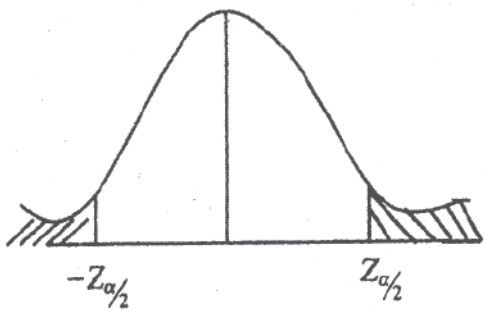
Power = 1 - Probability of type II error

$= P(x \in w \mid H_1)$

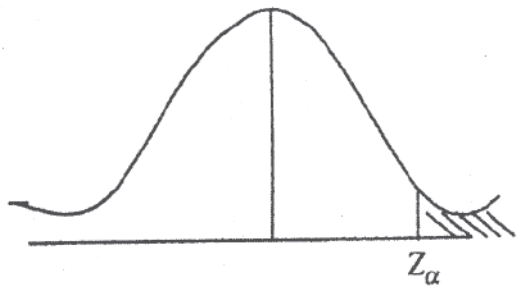
Two-tailed and one-tailed test

The specification of a critical region for a test depends upon the nature of the alternative hypothesis and the value of α. For example, $H_1 : \mu \neq \mu_0$, this implies that μ, may be less or greater than μ_0. Thus, the critical region is to be specified on "both tails of the curve with each part corresponding to half of the value of α. A test having critical region at both the tails of the probability curve is termed as a two tailed test.

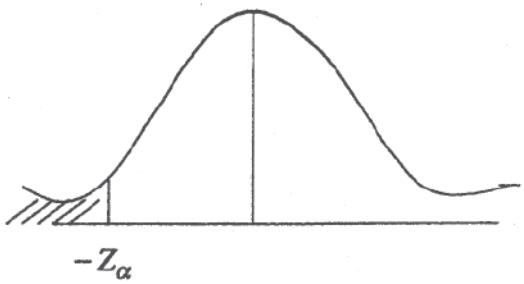
Further, if $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ the critical region is to be specified only at one -tail of the probability curve and the corresponding test is termed as a one-tailed test.



Critical region for two-tailed test.



Critical region for one-tailed (right tail) test



Critical region for one-tailed (left tail) test

6.2 Tests of Hypothesis Concerning Mean of Single Population

These tests can be divided into two broad categories depending upon whether σ, the population standard deviation, is known or not.

6.2.1 Test of Hypothesis Concerning Specified Mean (σ being known)

This test is applicable when the random sample $X_1, X_2, \ldots\ldots, X_n$ is drawn from a normal population with mean μ and standard deviation σ. We can consider the test in the folowing steps :

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the given problem. That is.

$H_0 : \mu = \mu_0$ (specified) against alternatives a) $H_1 : \mu > \mu_0$

or, b) $H_1 : \mu < \mu_0$

or, c) $H_1 : \mu \neq \mu_0$

Step II. (Test Statistic)

To test the above null hypothesis, we consider the appropriate test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim N(0,1) \text{ under the null hypothesis.}$$

Step III. (Computation)

$$Z_{cal} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \text{(say).}$$

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu > \mu_0$ at α level of significance if $Z_{Cal} > Z_\alpha$.

b) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu < \mu_0$

at α level of significance if $Z_{Cal} < -Z_{\alpha}$ and

c) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1: \mu \neq \mu_0$ at α level of significance if $| Z_{Cal} | > Z_{\alpha/2}$

Here Z_{α} denotes the upper α-point of a standard normal distribution. That is, $P(Z > Z_{\alpha}) = \alpha$, where Z is a standard normal variate.

Example 1. The mean breaking strength of cables supplied by a manufacturer is 1800 with a standard deviation 100. By a new technique in the manufacturing process, it is claimed that the breaking strengths of the cables have increased. In order to test the claim, a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 0.01 level of significance?

Solution: Let us assume that there is no significant change in mean breaking strength of 1800.

Null Hypothesis $H_0 : \mu = 1800$.

Against Alternating $H_1 : \mu > 1800$

Now, given that $\sigma = 100, n = 50, \bar{X} = 1850$.

The test statistic is $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \approx N(0,1)$ under the null hypothesis.

$$Z_{Cal} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} = \frac{\sqrt{50}(1850 - 1800)}{100} = 3.54$$

Since, Z_{Cal} $(=3.54) > Z_{0.01}$ $(=2.33)$, so the null hypothesis is rejected at 1% level of significance. Hence, the claim of breaking strengths increased is accepted.

Example 2. Construct 95% confidence interval for mean of a normal population.

Solution : Let $X_1, X_2 ... X_n$ be a random sample of size n from a normal population with mean μ and standard deviation σ.

We know that sampling distribution of $\bar{X}$ is normal with mean μ and standard

error $\frac{\sigma}{\sqrt{n}}$. Therefore, $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ will be a standard normal variate.

From the tables of areas under the standard normal curve, we can write

$P[-1.96 \leq Z \leq 1.96] = 0.95$ or $P\left[-1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96\right] = 0.95....(1)$

The inequality $-1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ can be written as

$-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu$ or $\mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$.....(2)

Similarly, from the inequality $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96$ we can write

$\mu \geq \bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$.....(3)

Combining (2) and (3), we get

$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$

Thus, we can write equation (1) as

$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$

This gives us 95% confidence interval for the parameter μ. The lower limit of μ is $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$ and the upper limit is $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$. The probability of μ lying between these limits is 0.95 and, therefore, this interval is also termed as 95% confidence interval for μ.

In a similar way, we can construct a 99% confidence interval for μ as

$$P\left(\bar{X} - 2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99.$$

Thus, the 99% confidence limits for μ are $\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

Remark : When σ is unknown and n < 30, we use t value instead of 1.96 or 2.58 and use s in place of σ.

6.2.2 Test of Hypothesis Concerning Specified Mean (σ being unknown)

This test is applicable when the random sample $X_1, X_2,..., X_n$ is drawn from a normal population with mean μ and standard deviation σ. We can consider the test in the following steps :

Step I. (Hypothesis Formulation)

The null hypothesis and the alternative hypothesis on the basis of the given problem are as follows :

$H_0 : \mu = \mu_0$ (specified) against alternatives a) $H_1 : \mu > \mu_0$

or, b) $H_1 : \mu < \mu_0$

or, c) $H_1 : \mu \neq \mu_0$

Step II. (Test Statistic)

To test the above null hypothesis, we consider the appropriate test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim t \text{ – distribution with } (n-1) \text{ degrees of freedom under}$$

null hypothesis, where $s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$.

Step III. (Computation)

Let the value of this statistic calculated from sample be denoted as

$$t_{Cal} = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{s} \text{ (say).}$$

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu > \mu_0$ at α level of significance if $t_{Cal} > t_{\alpha,\, n-1}$.

(b) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu < \mu_0$ at α level of significanxce if $t_{Cal} < -t_{\alpha,\, n-1}$ and

(c) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$ at α level of significance if $|t_{Cal}| > t_{\alpha/2,\, n-1}$.

Here $t_{\alpha,\, n-1}$ denotes the upper α-point of t-distribution with n–1 degrees of freedom. That is,

$P(t > t_{\alpha,\, n-1}) = \alpha$.

Note : When s is not known, we use its estimate computed from the given sample. Here, the nature of the sampling distribution of $\bar{X}$ would depend upon sample size n. There are the following two possibilities :

(i) If the parent population is normal and $n \leq 30$ (popularly known as small sample case), use t - test. The unbiased estimate of σ in this case is given by

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

(ii) If $n > 30$ (large sample case), use the standard normal test. The unbiased estimate of s in this case can be taken as $s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}}$, since the difference between n and n–1 is negligible for large values of n. Note that the parent population may or may not be normal in this case.

Example 3. A sample of 450 items is taken from a population whose standard deviation is 20. The mean of the sample is 30. Test whether the sample has come from

a population with mean 29. Also calculate the 95% confidence limits for the population mean.

Solution: Let us consider the hypothesis as follows:

Null Hypothesis $H_0 : \mu = 29$

Against Alternating $H_1 : \mu \neq 29$

Now, given that $n=450,\ \sigma=20,\ \bar{X}=30$.

The test statistic is $Z=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}=\dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}\sim N(0,1)$ under the null hypothesis.

$$Z_{Cal}=\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}=\frac{\sqrt{450}(30-29)}{20}=1.06$$

Since, $Z_{Cal}\ (=1.06) < Z_{0.05}\ (=1.96)$, so the null hypothesis is accepted at 5% level of significance. Hence, the sample has come from the population with mean 29.

2nd Part:

Again, 95% confidence limits for the population mean are

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow 30 \pm 1.96 \times \frac{20}{\sqrt{450}}$$

$$\Rightarrow 30 \pm 1.85 \ \Rightarrow 28.15, \text{ and } 31.85.$$

Example 4. A random sample of 10 boys had the following I.Q.'s :

70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.

Solution: Let us consider the Null hypothesis, that "The data are consistent with the assumption of a mean I.Q. of 100 in the population".

That is, Null Hypothesis $H_0 : \mu = 100$

Against Alternating $H_1 : \mu \neq 100$

The test statistic is: $t = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim t_{(n-1)}$

Where sample mean and variance are to be computed from the sample values of I.Q.'s.

X	**X–$\bar{x}$**	**(X–$\bar{x}$)**
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
$\sum X = 972$		1833.60

Hence, $n=10$, $\bar{X} = \frac{972}{10} = 97.2$, $s^2 = \frac{1833.60}{9} = 203.73$

$$|t| = \frac{\sqrt{10}(97.2 - 100)}{\sqrt{203.73}} = 0.62$$

And from the table $t_{0.05}(d.f = 9) = 2.262$

$|t|$ $(=0.62) < t_{0.05}$ $(=2.262)$

The null hypothesis is accepted at 5% level of significance and hence concluded that the data are consistent with the assumption of mean I.Q of 100 in the population.

Example 5. The heights of 10 males of a normal population are found to be 70, 67, 62, 67, 61,68, 70, 64, 65, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level assuming that for 9 degrees of freedom $P(t>1.83)=0.05$.

Solution:

Null Hypothesis $H_0 : \mu=64$

Against Alternative $H_1 : \mu>64$

The test statistic is: $t=\frac{\sqrt{n}(\bar{X}-\mu_0)}{s} \sim t_{(n-1)}$

From $P(t>1.83)=0.05$, the critical region is the interval $(1.83,\infty)$. To calculate the value of the test statistic, we consider the following table:

x_1	$y_1 = (x_1-64)$	y_i^2
70	6	36
67	3	9
62	-2	4
67	3	9
61	-3	9
68	4	16
70	6	36
64	0	0
65	1	1
66	2	4
$\sum x_i=660$	$\sum y_i=20$	$\sum y_i^2=124$

Therefore, $\bar{x}=\frac{\sum x_i}{n}=\frac{\sum y_i}{n}+64=\frac{20}{10}+64=66$

and sample variance $S^2=\frac{1}{n}\sum y_i^2-(\bar{y})^2=\frac{124}{10}-\left(\frac{20}{10}\right)^2=12.4-4=8.4$

Therefore, $s^2=\frac{n}{n-1}S^2=\frac{10}{9}\times 8.4=\frac{84}{9}$

Hence, $t=\frac{\sqrt{10}\,(66-64)}{\sqrt{84}/3}=\frac{6\sqrt{10}}{2\sqrt{21}}=\frac{3\sqrt{10}}{\sqrt{21}}=2.07$

Since, $t_{Cal}=2.07>1.83=t_{0.05}$. So, the null hypothesis is rejected at 5% level of significance. Hence, it is reasonable to believe that the average height is greater than 64 inches.

6.3 Test of Hypothesis Concerning Means of Two Populations

These tests can be divided into several categories depending upon whether σ_1 and σ_2, the population standard deviations are known or not. The populations may be dependent or independent.

6.3.1 Test of Hypothesis Concerning Equality of Two Means (σ_1 and σ_2 being known)

This test is applicable when two random samples of sizes n_1 or n_2 drawn from two independent normal populations with unknown means μ_1 and μ_2 and standard deviations $\sigma_1$and σ_2. We can consider the test in the following steps :

Step I. (Hypothesis Formulation)

We set up the null hypothesis and the alternative hypothesis on the basis of the above problem. That is,

$H_0 : \mu_1 = \mu_2$ against alternatives a) $H_1 : \mu_1 > \mu_2$

or, b) $H_1 : \mu_1 < \mu_2$

or, c) $H_1 : \mu_1 \neq \mu_2$

Step II. (Test Statistic)

To test the above null hypothesis, we consider the appropriate test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{*\sigma_1^2}{n_1} + \frac{*\sigma_2^2}{n_2}}} \sim N(0,1)$$ under the null hypothesis.

Step III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$$Z_{Cal} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ (say).

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 > \mu_2$ at α level of significance if $Z_{Cal} > Z_\alpha$.

(b) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 < \mu_2$ at α level of significance if $Z_{Cal} < -Z_\alpha$ and

(c) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ at a level of significance if $|Z_{Cal}| > Z_{\alpha/2}$.

Here Z_α denotes the upper α-point of a standard normal distribution. That is, $P(Z > Z_\alpha) = \alpha$, where Z is a standard normal variate.

6.3.2 Test of Hypothesis Concerning Equality of Two Means (σ_1 and σ_2 unknown but $\sigma_1 = \sigma_2$)

This test is applicable when two random samples of sizes n_1 or n_2 drawn from two independent normal populations with unknown means μ_1, μ_2 and standard deviations σ_1 and σ_2. The population standard deviations are assumed to be equal.

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the above

problem. That is,

$H_0 : \mu_1 = \mu_2$ against alternatives a) $H_1 : \mu_1 > \mu_2$

or, b) $H_1 : \mu_1 < \mu_2$

or, c) $H_1 : \mu_1 \neq \mu_2$

Step II. (Test Statistic)

To test the above null hypothesis, we consider the appropriate test statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_2 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$ under the null hypothesis, where

the pooled estimate of σ, denoted by s, is defined as

$$s = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

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

Step III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$$t_{Cal} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ (say).}$$

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 > \mu_2$ at α level of significance if $t_{Cal} > t_{\alpha,\ (n_1 + n_2 - 2)}$.

(b) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 < \mu_2$ at α level of significance if $t_{Cal} < -t_{\alpha;\,(n_1 + n_2 - 2)}$. and

(c) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ at α level of significance if $|t_{Cal}| > t_{\alpha/2;(n_1 + n_2 - 2)}$.

Here $t_{\alpha;(n_1 + n_2 - 2)}$. denotes the upper α-point of a t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. That is, $P(t > t_{\alpha;(n_1 + n_2 - 2)}) = \alpha$.

Note : If we consider two random samples of sizes n_1 or n_2 drawn from two independent normal populations with unknown means μ_1, μ_2 and standard deviations σ_1, σ_2 respectively, then

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Case I : If σ_1 and σ_2 are known, we use the standard normal test.

(a) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ (two-tailed test) the test statistic is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0.$$

This value is compared with 1.96 (2.58) for 5% (1%) level of significance.

(b) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ (one-tailed test), the test statistic is $Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim$ N (0,1) under H_0. and the critical value of 5% (1%) level of significance is 1.645 (2.33).

(c) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ (one-tailed test), the test statistic Z_{Cal} is same as in (b) above, however, the critical value for 5% (or 1%) level of significance is – 1.645 (or –2.33).

Case II : If σ_1 and σ_2 are not known, their esimates based on samples are used. This category of tests can be further divided into two sub-groups.

Small sample tests (when either n_1 or n_2 or both are less than or equal to 30). For this test $H_0 : \mu_1 = \mu_2$, we use t-test. The respective estimates of σ_1 and σ_2 are given by

$$S_1 = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1}} = S_1\sqrt{\frac{n_1}{n_1 - 1}} \quad \text{and} \quad S_2 = \sqrt{\frac{\sum\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1}} = S_2\sqrt{\frac{n_2}{n_2 - 1}}$$

This test is more restrictive because it is based on the assumption that the two samples are drawn from independent normal populations with equal standard deviations. i.e. $\sigma_1 = \sigma_2 = \sigma$ (say). The pooled estimate of σ, denoted by s, is defined as

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

$$= \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{n_1 + n_2 - 2}}$$

(a) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ (two-tailed), the test statistic is

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$ which follows the t-distribution with $(n_1 + n_2 - 2)$ d.f.

(b) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ (one-tailed test), the test statistic is $t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$.

(c) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ (one-tailed test), the test statistic, i.e. t is same as in (b). This value is compared with the negative t value.

2. Large Sample test (when each of n_1 and n_2 is greater than 30)

In this case s_1 and s_2 are estimated by their respective sample standard deviations S_1 and S_2. The test statistic for two and one-tailed test is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$ and the remaining procedure is same as above.

Remark : 100 (1–α)% confidence limits for $h_1 - h_2$ are given by $\left(\bar{X}_1 - \bar{X}_2\right) \pm Z_{\alpha/2}$ $S.E._{\bar{X}_1 - \bar{X}_2}$. If the two samples are drawn from populations with same standard deviation, i.e. $\sigma_1 = \sigma_1 = \sigma$ (say), then $S.E._{\bar{X}_1 - \bar{X}_2} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case I and $S.E._{\bar{X}_1 - \bar{X}_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case II. For large sample tests, s can also be estimated by S as

$$S = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}}.$$

Example 6. In a random sample of size 500, the mean is found to be 20. In another independent sample of size 400, the mean is 15. Could the samples have been drawn from the same population with standard deviation 4? (Given that $Z_\alpha = 2.58$ at 1% level of significance)

Solution:

Here $\bar{X}_1 = 20, n_1 = 500$; $\bar{X}_2 = 15, n_2 = 400$ *and* $\sigma = 4$.

Null Hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

To test the above null hypothesis the test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20 - 15}{4\sqrt{\frac{1}{500} + \frac{1}{400}}} = 18.6$$

$\therefore$ Z_{Cal} $(=18.6) > Z_{0.01}$ $(=2.58)$

Hence, Null hypothesis H_0 is rejected i.e., the samples have not been drawn from the same population.

Example 7. Two samples are drawn from two normal populations with same standard deviation. First sample is of size 100, mean 61 and s.d 4 and another sample of size 200 is having mean 63 and s.d 6. Test at 5% level of significance about the significance of difference between the sample means.

Solution:

Here $n_1 = 100, \bar{X}_1 = 61, s_1 = 4$; $n_2 = 200, \bar{X}_2 = 63$ *and* $s_2 = 6$.

Null Hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

To test the above null hypothesis the test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{61 - 63}{\sqrt{\frac{4^2}{100} + \frac{6^2}{200}}} = -3.02$$

$\therefore$ $|Z_{Cal}|$ $(=3.02) > Z_{0.05}$ $(=1.96)$

Hence, Null hypothesis H_0 is rejected i.e., the two normal populations from which the samples are drawn, may not have the same mean.

6.3.3 Test of Hypothesis Concerning Equality of Two Means (σ_1 and σ_2 unknown for a bivariate population)

Suppose we have a random sample of n pairs of observations from a bivariate normal population with unknown means μ_1, μ_2 and standard deviations σ_1 and σ_2. The population standard deviations are not assumed to be equal. Suppose the paired data are available (X_1, Y_i), i = 1, 2, 3, ..., n, j = 1, 2, 3 .....m_1

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the above problem. That is, $H_0 : \mu_1 = \mu_2$ against alternatives

a) $H_1 : \mu_1 > \mu_2$

or b) $H_1 : \mu_1 < \mu_2$

or c) $H_1 : \mu_1 \neq \mu_2$

Step II. (Test Statistic)

Let $U_i = X_i - Y_i$, difference in the values of X and Y for the i-th pair, i = 1, 2, 3, ..., n. To test the above null hypothesis we consider the appropriate test statistic $t = \frac{\sqrt{n}\bar{U}}{S_U}$ which follows the t-distribution with (n–1) degrees of freedom under the null hypothesis, where $S_U = \sqrt{\frac{\Sigma\left(U_i - \bar{U}\right)^2}{n-1}}$.

Step III. (Computation)

Let the value of this statistic calculated from the sample be denoted as $t_{Cal} = \frac{\sqrt{n}\bar{U}}{S_U}$ (say).

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 > \mu_2$ at α level of significance if $t_{Cal} > t_{\alpha,n-1}$.

(b) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 < \mu_2$ at α level of significance if $t_{Cal} < -t_{\alpha,\ n-1}$ and

(c) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1\ \mu_2$ at α level of significance if $|t_{Cal}| > t_{\alpha/2,\ n-1}$.

Here $t_{\alpha;\ n-1}$ denotes the upper α-point of a t-distribution with n–1 degrees of freedom. That is, $P(t > t_{\alpha,\ n-1})=\alpha$.

Note : The above test is known as Paired t-test which may be viewed as Student's t-test with n – 1 degrees of freedom.

6.4 Test of Hypothesis for Proportion

6.4.1 Test of Hypothesis for Specified Population Proportion

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the given problem. That is, the null hypothesis to be tested is

$H_0 : \pi = \pi_0$ against alternatives a) $H_1 : \pi > \pi_0$

or, b) $H_1 : \pi < \pi_0$

or, c) $H_1 : \pi \neq \pi_0$

Stem II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = (p - \pi_0)\sqrt{\frac{n}{\pi_0(1-\pi_0)}} \sim N(0,1)$$ under the null hypothesis for sufficiently large n.

Step III. (Computation)

Let the value of this statistic calculated from sample be denoted as

$$Z_{Cal} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \text{(say).}$$

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \pi = \pi_0$ against the alternative $H_1 : \pi > \pi_0$ at α level of significance if $Z_{Cal} > Z_\alpha$.

(b) Reject the null hypothesis $H_0 : \pi = \pi_0$ against the alternative $H_1 : \pi < \pi_0$ at α level of significance if $Z_{Cal} < -Z_\alpha$ and

(c) Reject the null hypothesis $H_0 : \pi = \pi_0$ against the alternative $H_1 : \pi \neq \pi_0$ at α level of significance if $|Z_{Cal}| > Z_{\alpha/2}$.

Here Z_α denotes the upper α-point of the standard normal distribution. That is, $P(Z>Z_\alpha) = \alpha$, where Z is a standard normal variate.

Remark : The 100 $(1-\alpha)$% confidence limits for p are $\pi \pm Z_{\alpha/2}$S.E.(p).

Example 8 : A certain controlled process produces 15 percent defective items. A supplier of a basic raw material claims that the use of his material would reduce the fraction of defective. On making a production trial run with the new material, it was found that from an output of 400 units 52 were defective. Would you accept the supplier's claim?

Solution. Let X be the random variable denoting the number of defectives in a sample of size n. We assume that X follows the binomial distribution with parameters n and π.

Here we have to test the null hypothesis

$H_0 : \pi = 0.15$ against the alternative $H_1 : \pi < 0.15$.

$p = X/n = 52 / 400 = 0.13$

$$Z_{Cal} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.13 - 0.15}{\sqrt{\frac{0.15 \times 0.85}{400}}} = -1.12.$$

$Z_{0.05} = -1.645$.

i.e. at 5% level of significance the supplier's claim that the new material will reduce the fraction of defective is rejected at 5% level of significance.

Example 9 : A random sample of 100 items taken from a large batch of articles contains 5% defective items, (a) Set up 96% confidence limit for the proportion of defective items in a batch, (b) If the batch contains 2, 697 items, set up the 95% confidence limits for the proportion of defective items.

Solution : Here n = 100.

p = proportion of defectives in the sample = 5/100 = 0.05

(a) Here estimate of S.E. of p is given by

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05 \times .95}{100}} = 0.02179 \cong 0.022$$

From the table of the normal distribution $Z_{0.02} = 2.05$.

Hence 96% confide nee limits for the population proportion of defectives are :

$p \pm Z_{0.02}$ S.E.(p) = $0.05 \pm 2.05 \times 0.022 = (0.005, 0.095)$

(b) If N = 2669, then 95% confidence limits for p are given by

$$P \pm Z_{0.025}\ SE(p) = p \pm 1.96 \times \sqrt{\frac{(N-n)pq}{N(n-1)}}.$$

$$\left[\sqrt{\frac{pq}{n}\frac{(N-n)}{N-1}} \text{ is biased but } \sqrt{\frac{pq}{n-1}\frac{(N-n)}{N}} \text{ is unbiased}\right]$$

$$= 0.050 \pm 1.96 \times \sqrt{\frac{(2669-100)}{2669} \times \frac{0.05 \times 0.95}{99}}$$

$= 0.050 \pm 1.96 \times 0.0215 = 0.050 \pm 0.042$

$= (0.008, 0.092)$

6.4.2 Test of Hypothesis Concerning Equality of Proportions

Setp I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the given problem. That is, the null hypothesis to be tested is

$H_0 : \pi_1 = \pi_2$ against the alternatives a) $H_1 : \pi_1 > \pi_2$

or, b) $H_1 : \pi_1 < \pi_2$

or, c) $H_1 : \pi_1 \neq \pi_2$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$Z = \frac{(p_1 - p_2)}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{\pi(1-\pi)(n_1 + n_2)}} \sim N(0, 1)$$ under the null

hypothesis for sufficiently large n_1 and n_2 under the assumption that $\pi_1 = \pi_2 = \pi$, where π is known. Often population proportion π is unknown and it is estimated on the basis of samples. The pooled estimate of π, denoted by p, is given by $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

Thus, the test statistic becomes $Z = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{p(1-p)(n_1 + n_2)}}$

Setp III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$$Z_{Cal} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{p(1-p)(n_1 + n_2)}}$$ (say)

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \pi_1 = \pi_2$ against the alternative $H_1 : \pi_1 > \pi_2$ at α level of significance if $Z_{Cal} > Z_\alpha$.

(b) Reject the null hypothesis $H_0 : \pi_1 = \pi_2$ against the alternative $H_1 : \pi_1 < \pi_2$ at α level of significance if $Z_{Cal} < - Z_\alpha$ and

(c) Reject the null hypothesis $H_0 : \pi_1 = \pi_2$ against the alternative $H_1 : \pi_1 \neq \pi_2$ at α level of significance if $| Z_{Cal} | > Z_\alpha$.

Here Z_α denotes the upper α - point of standard normal distribution. That is. $P(Z > Z_\alpha) = \alpha$ where Z is a standard normal variate.

Remark : The 100 $(1-\alpha)$% confidence limits for $\pi_1 - \pi_2$ are $(p_1 - p_2) \pm Z_{\alpha/2}$ S E $(p_1 - p_2)$.

Example 10. A survey of television audiance in a big city revealed that a particular programme was liked by 50 out of 200 males and 80 out of 250 females. Test the hypothesis that whether there is a real difference of opinion about the programme between males and females.

Solution : Let π_1 and π_2 be the proportion of males and females who liked the particular television prgaramme. The null hypothesis to be tested is

$H_0 : \pi_1 = \pi_2$ against alternative $H_1 : \pi_1 \neq \pi_2$.

To test the above null hypothesis we condsider the appropriate test statistic

$$Z = \frac{(p_1 - p_2)}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = (p_1 - p_2\sqrt{\frac{n_1 n_2}{\pi(1-\pi)(n_1 + n_2)}} \sim N(0,1) \text{ under the null}$$

hypothesis for large n_1 and n_2 under the assumption that $\pi_1 = \pi_2 = \pi$ where π is known. Often population proportion π is unknown and it is estimated on the basis of samples.

The pooled estimate of π denoted by p is given by $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Here $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{50 + 80}{200 + 250} = \frac{13}{45}$

$$\text{So, } Z_{Cal} = \frac{\frac{80}{250} - \frac{50}{200}}{\sqrt{pq\left(\frac{1}{250} + \frac{1}{200}\right)}} = \frac{0.32 - 0.25}{0.04305} = 0.958.$$

Since the computed value of Z (0.958) is less than the tabulated value of Z (1.96) at 5% level of isgnificance, so the null hypothesis is accepted. That is, there is no real difference of opinion about the programme between males and females.

Example 11: Obtain the 95% confidence limits for the proportion of success in a binomial population.

Solution : Let the parameter π denote the proportion of successes in population. Further, p denote the proportion of successes in n ($\neq$50) trials. We know that the sampling distribution of p will be approimately normal with mean p and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$.

Since π is not known, therefore, its estimator p is used in the estimation of standard error of p, i.e. S.E.(p) = $\sqrt{\frac{p(1-p)}{n}}$.

Thus, the 95% confidence interval for p is given by

$$p\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95.$$

This gives the 95% confidence limits as $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$.

Example 12 : In a newspaper article of 1600 words in Hindi, 64% of the words hold good, estimate the confidence limits for the proportion of Sanskrit words in the writer's vocabulary.

Solution : Let π be the proportion of Sanskrit words in the writer's vocabulary. Corresponding proportion in the sample is given as p = 0.64

$$\therefore \text{S.E.}(p) = \sqrt{\frac{0.64 \times 0.36}{1600}} = \frac{0.48}{40} = 0.012.$$

We known that almost whole of the distribution lies between 3σ limits. Therefore, the confidence interval is given by

$P[p - 3\text{S.E.}(p) \leq \pi \leq p + 3\text{S.E.}(p)] = 0.9973$

Thus, the 99% confidence limits are 0.609 (=0.64 – 2.58 × 0.012) and 0.671 (=0.64 + 2.58 × 0.012) respectively.

Hence, the proportion of Sanskrit words in the writer's vocabulary are between 60.9% and 67.1%.

6.5 Test of Hypothesis Concerning Population Standard Deviation

Step I. (Hypothesis Formulation)

We set up the null hypothesis and altermative hypothesis on the basis of the given problem. That is, the null hpothesis to be tested is

$H_0 : \sigma = \sigma_0$ against the alternatives a) $H_1 : \sigma > \sigma_0$

or, b) $H_1 : \sigma < \sigma_0$

or, c) $H_1 : \sigma \neq \sigma_0$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$\chi^2 = \frac{\sum(X_i - \bar{X})^2}{\sigma_0^2} = \frac{nS^2}{\sigma_0^2}$ follows a χ^2 - variate with (n – 1) degrees of freedom under the null hypothesis.

Setp III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$$\chi^2_{\text{Cal}} = \frac{\sum(X_i - \bar{X})^2}{\sigma_0^2} = \frac{nS^2}{\sigma_0^2} \text{(say)}.$$

Step IV. (Conclusion)

(c) Reject the null hypothesis $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma > \sigma_0$ at α level of significance if $\chi^2_{Cal} > \chi^2_{\alpha;(n-1)}$.

(d) Reject the null hypothesis $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma < \sigma_0$ at α level of significance if $\chi^2_{Cal} < \chi^2_{1-\alpha;(n-1)}$ and

(e) Reject the null hypothesis $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma \neq \sigma_0$ at α level of significance if $\chi^2_{Cal} > \chi^2_{\frac{\alpha}{2};(n-1)}$ or $\chi^2_{Cal} > \chi^2_{\frac{1-\alpha}{2};(n-1)}$.

Here $\chi^2_{\alpha;(n-1)}$ denotes the uper α - point of a χ^2 - variate with (n – 1) degrees of fredom. That is, $P\left(\chi^2 > \chi^2_{\alpha;(n-1)}\right) = \alpha.$

Note : It can be shown that for large sample (n > 30), the sampling distribution of S is approximatelity normal with mean α and standard error $\frac{\sigma}{\sqrt{2n}}$. Thus $Z = \frac{(S-\alpha)\sqrt{2n}}{\sigma}$ ~ N(0,1) for sufficiently large value of n.

Alternatively, using Fisher's approximation, we can say that when n > 30 the statistic $\sqrt{2\chi^2}$ follows a normal distribution with mean $\sqrt{2n-1}$ and standard error unity. Thus $Z = \sqrt{2\chi^2} - \sqrt{2n-1}$ can be taken as a standard normal variate for sufficiently large values of n.

Example 13. The life time of certain batteries are supposed to have the variance 5000 hrs. Test at 5% level of significance the hypothesis that $\sigma^2 = 5000$ hrs against the alternating hypothesis $\sigma^2 > 5000$, if 30 of these batteries had an unbiased estimate of sample variance 7500 hrs.

$$\left[P(\chi^2 > 42.77) = 0.05 \text{ for } 29 \text{ d.f} \right]$$

Solution:

Null Hypothesis $H_0 : \sigma^2 = 5000 = \sigma_0^2$

Against Alternating $H_1 : \sigma^2 > 5000 = \sigma_0^2$

Sample size n = 30 and variance s^2 = 7500.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{29 \times 7500}{5000} = 43.5$$

$$\chi^2 (=43.5) > \chi^2_{0.05,29} (=42.557)$$

Therefore, Null hypothesis H_0 is rejected.

Hence, we cannot say that $s^2 \neq 5000$.

6.6 Test of Hypothesis Concerning the Equality of Standard Deviations

Step I. (Hypothesis Formulation)

We set up the null hypothesis and the alternative hypothesis on the basis of the given problem. That is, the null hypothesis to be tested is

$H_0 : \sigma_1 = \sigma_2$ against alternative $H_1 : \sigma_1 > \sigma_2$.

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic $F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$ which would become $\frac{s_1^2}{s_2^2}$ and under H_0 it follows the F – distribution with v_1 (= $n_1 - 1$) and v_2 (= $n_2 - 1$) degrees of freedom.

Setp III. (Computation)

Let the value of this statistic calculated from the sample be denoted as $F_{Cal} = \frac{s_1^2}{s_2^2}$ (Say).

Setp IV. (Conclusion)

Reject the null hypothesis $H_0 : \sigma_1 = \sigma_2$ against the alternative $H_1 : \sigma_1 > \sigma_2$ at α level of significance if $F_{Cal} > F_{\alpha;(n_1-1),\,(n_2-1)}$.

Here $F_{\alpha;(n_i-1)(n_2-1)}$ denotes the upper α - point of an F - variate with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. That is, $P(F > F_{\alpha;(n_1-1),\,(n_2-1)} = \alpha$.

Remarks:

1. We can write $S_1^2 = \frac{1}{n_1 - 1}\Sigma(X_{1i} - \bar{X}_1)^2 = \frac{n_1}{n_1 - 1}s_1^2 = \frac{1}{n_1 - 1}\left(\Sigma X_{1i}^2 - \frac{(\Sigma X_{1i})^2}{n_1}\right)$

   and $S_2^2 = \frac{1}{n_2 - 1}\Sigma(X_{2i} - \bar{X}_1)^2 = \frac{n_2}{n_2 - 1}s_2^2 = \frac{1}{n_2 - 1}\left(\Sigma X_{2i}^2 - \frac{(\Sigma X_{2i})^2}{n_1}\right)$.

2. In the variance ratio $F = \frac{S_1^2}{S_2^2}$, we take, by convenion, the largest of the two sample variances as S_1^2. Thus, this test is always a one-tailed test with critical region at the right hand tail of the F - distribution.

3. The 100 $(1 - \alpha)$% confidence limits for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$ are given by

$$P\left(\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\alpha/2}}\right) = 1 - \alpha.$$

6.7 Frequeny Chi-square (Pearsonian χ^2)

6.7.1 Test for Goodness of Fit

The use of chi-square (χ^2) test was first devised by Karl Pearson to decide whether the observations are in good agreement with a hypothetical distribution i.e., whether the sample may be supposed to have come from a specified population. The

observed values (f_o) for different classes are compared with expected values (f_e) forming the test statistic.

$$\therefore \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \sim \chi^2_{k-1}.$$

This is called the goodness of fit Chi-square or Pearsonian Chi-square or freqhency Chi-square distribution with k-1 degrees of freedom under the null hypothesis where k is the number of classes. We reject the null hypothesis if the observed value of the statistic exceeds the tabulated value of χ^2 at a particular level.

Example 14. The following data represent the number of accident happened in different days of a week.

Day :	Mon	Tues	Wed	Thurs	Fri	Sat
No. of Accidents :	15	19	13	12	16	15

Test whether the accidents are uniformly distributed over the week.

Solution:

Let the null hypothesis is that the accidents occur uniformly over the week.

Total number of accidents = 90.

Thus, the expected number of accidents on any day of the week $=\frac{90}{6}=15$

f_e :	15	19	13	12	16	15
f_o :	15	15	15	15	15	15

$$\therefore \chi^2=\sum \frac{(f_O-f_e)^2}{f_e}=\frac{1}{15}(0+16+4+9+1+0)=2$$

The degree of freedom = 6-1=5.

Now, from χ^2 table $\chi^2_{0.05}(d.f=5)=11.07$.

Since the observed value of $\chi^2(=2)$ is less the tabulated value $\chi^2_{0.05}(d.f=5)=11.07$, null hypothesis is accepted at 5% level of significance and hence concluded that the accidents occurred uniformly over the week.

Example 15. A survey of 320 families with 5 children each revealed the following distribution:

No. of boys:	0	1	2	3	4	5
No. of girls:	5	4	3	2	1	0
No. of families:	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable?

Solution:

Let p = P (Male birth) = $\frac{1}{2} = P(Female\ Birth) = q$

Let the null hypothesis be that the male and female births are equally probable.

Let P(r) = Probability of r male birth in a family of 5 $= {}^5C_r\ p^r\ q^{n-r} = {}^5C_r \left(\frac{1}{2}\right)^5$

Therefore, the frequency of r male birth $f(r) = N \times P(r) = 320 \times {}^5C_r \left(\frac{1}{2}\right)^5 = 10.{}^5C_r$

(Where, N = Total number of families =320)

Thus, $f(0) = 10.{}^5C_0 = 10$

$f(1) = 10.{}^5C_1 = 50$

$f(2) = 10.{}^5C_2 = 100$

$f(3) = 10.{}^5C_3 = 100$

$f(4) = 10.{}^5C_4 = 50$

$f(5) = 10.{}^5C_5 = 10$

Observed Frequency (f_O)	Expected Frequency (f_e)	$(f_O - f_e)^2$	$(f_O - f_e)^2 / f_e$
14	10	16	1.60
56	50	36	0.72
110	100	100	1.00
88	100	144	1.44
40	50	100	2.00
12	10	4	0.40
$\sum f_O = 320$	$\sum f_e = 320$		Total: 7.16

$$\therefore \chi^2 = \sum \frac{(f_O - f_e)^2}{f_e} = 7.16$$

Now, the tabulated value (at 5% level of significance) of $\chi^2_{0.05}(d.f=5)=11.07$.

Since the observed value of $\chi^2 (=7.16)$ is less the tabulated value $\chi^2_{0.05}(d.f=5)=11.07$, null hypothesis is accepted at 5% level of significance and hence concluded that the null hypothesis of male and female birth equally probable is accepted.

6.7.1 Test for Independence of Attributes

When observations are classified according to two attributes and arranged in a two-way table, the display is put in terms of a contingeny table.

Two-Way Contingency Table

Attribute B	**Attribute A**	**Total**
	A_1 A_2 ..... An	
B_1	O_{11} O_{12} ..... O_{1n}	R_1
B_2	O_{21} O_{22} ..... O_{2n}	R_1
*	*	*
*	*	*
B_m	O_{ml} O_{m2} ..... O_{mn}	R_m -
Total	**C_l C_2 ..... C_n**	**N**

Here it may be noted that the attributes A and B have been classified into mutually exclusive categories. The value O_{ij} represents the frequency of the observation corresponding to the i-th row and j-th column. The expected frequency

E_{ij} is given by $E_{ij} = \frac{R_i}{N} \times \frac{C_j}{N} \times N = \frac{R_i \times C_j}{N}, i = 1, 2, 3, ..., n;\ j = 1,2,.....,m.$

Here the null hypothesis H_0 : A and B are independent

against the alternative H_1 : A and B are dependent

The test statistic under the null hypothesis for the test is

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} \sim \chi^2_{(m-1)(n-1)}.$$

We reject the null hypothesis at α% level of significance if observed

$\chi^2 > \chi^2_{\alpha,(m-1)(n-1)}$.

Example 16 : In a recent diet survey, the following results were obtained in an Indian city :

No. of families	**Hindus**	**Muslims**	**Total**
Tea takers	1236	164	1400
Non-tea takers	564	36	600
Total	**1800**	**200**	**2000**

Discuss whether there is any sigificant difference between the two communities in the matter of taking tea. Use 5% level of significance.

Solution : The null hypothesis that is to be tested can be written as H_0 : There is no difference between the two communities in the matter of taking tea.

Usine the direct formula, we have $\chi^2 = \frac{2000(1236 \times 36 - 164 \times 564)^2}{1400 \times 1800 \times 200 \times 600} = 15.24.$

The value of χ^2 from the table for 1 d.f. and at 5% level of significance is 3.84. Since the calculated value is greater than the tabulated value, H_0 is rejected. That is, there is a significant difference between the two communities in the matter of taking tea at 5% level of significance.

Example 17 : A certain drug is claimed to be effective in curing clods. In an expenriment on 500 persons with clods, half of them were given the drug. The

patients' reaction to the treatment are recorded in the following table.

Treatment	Helped	Reaction	No efect	Total
Drug	150	30	70	250
Sugar Pills	130	40	80	250
Total	**280**	**70**	**150**	**500**

On the basis of the data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills? (Given $\chi^2_{005,\ 2} = 5.99$).

Solution : Let us take the null hypothesis that there is no significant difference in the effect of the drug and sugar pills.

The contingency table is of size 2 × 3, the degree of freedom would be (2 – 1) (3 – 1) = 2. The expected frequencies can be calculated in the follwoing way :

$E_{11} = \frac{280 \times 250}{500} = 140$, and so on.

Contingency table for expected frequencies is as follows :

Treatment	Helped	Reaction	No efect	Total
Drug	140	35	75	250
Sugar Pills	140	35	75	250
Total	**280**	**70**	**150**	**500**

Arranging the observed and the expected frequencies in the following table we calculate the value of χ^2 test statistic.

Cell (i.j)	Observed frequency (O)	Expected frequency (E)	$(O - E)^2/E$
1,1	150	140	0.714
2,1	130	140	0.714
1,2	30	35	0.714
2,2	40	35	0.714
1,3	70	75	0.333
1,3	80	75	0.333
Total	**500**	**500**	**3.522**

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 3.522.$$

Since, the observed $\chi^2 < \chi^2_{0.05,2}$, therefore, we accept the null hypothesis at

5% level and conclude that there is no significant difference in the effect of the drug and suggar pills.

6.8 Summary

In this unit, we have presented the concepts of estimation and tests of hypotheses. In contrast, to test of hypothesis attempts to answer, what is the numerical value of a parameter and whether there is enough evidence to support the alternative hypothesis. We have discussed how to tests the hypotheses about population mean, standard deviation and population proportion for a single population and also extended for bivariate population. The rejection or acceptance of null (alternative) hypothesis against the alternative (null) hypothesis explained with number of examples.

6.9 Self-Assessment Questions

1. What is test of significance? Explain the procedure generally followed in testing of hypothesis.
2. Distinguish between (a) critical region and acceptance regions, (b) null hypothesis and alternative hypothesis, (c) one-tailed and two-tailed test, (d) type I error and type II error.
3. Explain clearly the procedure of testing hypothesis. Also point out the assumptions in hypothesis testing in large samples.
4. How does small sampling theory differ from large sampling theory?
5. Explain the following terms : test statistic, level of significance, confidence level and power of a test.
6. Give some important applications of't' test and explain how it helps in business decision making.
7. Discuss the F-test for testing the equality of two variances.
8. What is chi-square test? Explain its important uses with the help of examples.

9. A sample of 25 male students is found to have a mean height of 171.38cm. Can it be reas: r.ar ; :e carded as a sample from a population with a mean height of 171.17 cms and a standard deviation of 3.30 cms?

10. A toothpase company conducted a survey and found that it could sell only 60 tubes on an average per month per shop. Immediately, the company advertised heavily in sevenl media and after 3 months again conducted a survey and f: _r.c v.e saies was S3 tubes with a standard deviation of 10 tubes in a sample of 20 shops. Can it be concluded that the advertiesement is effective? ($t_{0.05;19}$ = 1.729, $t_{0.01;\ 19}$ = 2.861)

11. In a survey at a super market, the following number of people were observed purchasing different brands of coffee :

A	**B**	**C**	**D**	**E**
74	53	81	70	82

Do these data support the hypothesis that the population of coffee buyers prefer each of the five brands equally?

Hint : $\chi^2_4 = \dfrac{2^2 + 19^2 + 9^2 + 2^2 + 10^2}{72} = 7.64$

12. A sample of 540 households was selected to study the occupational pattern of the father and the son. The number of households obtained has been tabulated below. Test the hypothesis that the son's occupation is independent of the father's occupation.

Father's occupation

		Son's occupation			
		F	B	S	M
	Farming	24	97	62	58
Father's occupation	Business	22	28	30	41
	Services	32	10	11	20
	Miscellaneous	38	25	14	28

($\chi^2_{9;\ 0.05}$ = 16.919. $\chi^2_{9;\ 0.01}$ = 21.666).

13. An automobile manufacturing firm is bringing out a new model. In order to map out its advertising campaign, it wants to determine whether the model will appeal most to a particular age group or equally to all age groups. The firm conducted a survey and the results are summarised below :

	Age group			
	Below 20	**20-39**	**40-59**	**60 and above**
Liked	146	78	48	28
Disliked	54	52	32	62

What conclusion would you draw from the above data?

$\chi^3_{3;\ 0.05} = 7.815$, $\chi^2_{3;0.01} = 11.341$.

14. A man buys 15 electric bulbs of 'Philips' make and another 10 of the 'GE' make. He finds that the Philips bulbs give an average life of 1200 hours with S.D. of 60 hours and the GE bulbs give an average of 1242 hours with an S.D. of 80 hours. Is there a significant difference between the two makes?

($t_{23;0.025} = 2.069$, $t_{23;\ 0.005} = 2.806$)

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

Shops :	A	B	C	D	E	F
Before campaign :	53	28	31	48	50	42
After campaign :	58	29	30	55	56	45

Can the campaign be judged to be a success? $t_{5\ ;\ 0.05} = 2.015$)

16. Two types of scooters manufactured in India are tested for petrol mileage. One group consisting of 12 scooters with average mileage of 44 km/It and of standard deviation of 2 km while the other group consisting of 10 scooters with an average mileage of 50 km/ It and standard deviation of 1.5 km of petrol. Test whether statistically there exists a significant difference in the petrol consumption of two types of scooters. ($t_{20;\ 0\ 025} = 2.086$, $t_{20;0005} = 2.845$)

17. Two laboratories A and B carry out independent estimates of fat content in ice-cream made by a firm. A sample is taken from each population, halved and the separate halves sent to the two laboratories. The fat content obtained by the

laboratories is recorded below :

Batch No :	1	2	3	4	5	6	7	8	9	10
Lab A :	3	5	7	3	8	6	9	3	7	8
Lab B :	9	8	8	4	7	7	9	6	6	6

Is there a significant difference between the mean fat content obtained by the two laboratories A and B? ($t_{9\ 0025}$ = 2.262, $t_{9.0005}$ = 3.250)

18. A company making a brand of detergent and toilet soap wanted to compare the expenses incurred on sales promotion for these two products. The data for the preceding > ear were retrieved from the books of accounts of these two products and they are reproduced below :

Expenditure in Rs. Thousand

Months Product	1	2	3	4	5	6	7	8	9	10	11	12
Toilet soap	55	80	50	60	50	60	70	45	50	60	60	70
Detergent	50	25	70	45	60	55	45	60	55	55	45	35

Further, suppose that both the products had offered equal amount of profitability and turnover. Then verify whether the above sales promotion expenditure are justifiable or not. ($t_{0.025;\ 11}$ = 2.20, $t_{0.005;11}$ = 3.11).

19. The following data were obtained from a test in a laboratory.

Method	sample size	sample variance
A	10	1296
B	15	784

Test whether there is any difference in the variances of two methods at 5% level.

($F_{9,14;0.05}$ = 2.65, $F_{9,14;001}$ = 4.03)

20. A random sample of 15 observations gave an unbiased estimator s^2 = 12.63 of the population variance σ^2. May the sample be reasonably regarded as drawn from a normal population with variance 8? Test at 5% level of significance.

($\chi^2_{0.05;\ 14}$ = 23.68, $\chi^2_{0.01;14}$ = 29.14)

21. A stock broker claims that he can predict with 80% accuracy whether the values of a stock will rise or fall during the coming month. As a test he predicts the outcome of 40 stocks and is correct in 28 of the predictions. Does the evidence support the stock broker's claim?

22. 500 units from a factory are inspected and 12 are found to be defective. Similarly, 800 units from another factory are inspected and 17 are found to be defective. Can it be concluded that production in the second factory is better than in the first?

23. Determine the sample size for estimating the true weight of tea containers from (i) a large number of containers and (ii) from 1000 containers so that the estimate should be within 10 gms of the true average weight. Variance is 40 gms (on the basis of past record).

Hint:

$E = |\bar{X}-\mu| = 10$ $\sigma^2 = 40$

$$n = \frac{\sigma^2 z^2_{\alpha/2}}{E^2}$$ for SRSWR

$$n = \frac{N\left(\sigma^2 z^2_{\alpha/2}\right)}{NE^2\sigma^2 z^2_{\alpha/2}}$$ for SRSWR

For proportion

$$n = \frac{p(1-p)z^2_{a/2}}{E^2}$$ for SRSWR

$$n = \frac{p(1-p)z^2_{a/2}}{NE^2 + p(1-p)z^2_{a/2}}$$ for SRSWR

Unit 7 ❑ Analysis of Variance

Structure

7.0 Objectives

After studying the present unit, you will be able to (i) understand the necessity for analyzing data from more than two samples; (ii) understand the basic models to analysis of variance; (iii) explain the method of one-way as well as two-way analysis of variance; and (iv) construct the table for one-way and two-way analysis of variance.

7.1 Introduction

In case of two independent normal populations, t-test (Student's distribution) or z-test is used to compare any significant difference between the means of those two populations exists or not. This t-statistic is fails if the number of populations is more than two. If the number of populations is more than two, F-distribution is used for testing the equality of means of all the populations comparing the sample variances.In this situation, Analysis of variance (ANOVA) technique plays an important role for more than two populations. In other words, ANOVA is a technique used to test the null hypothesis that the means of three or more populations are equal. Using this technique, one can draw inferences whether the samples have been drawn from

populations having the same mean. The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. This technique of testing null hypothesis[1] (all the population means are equal) against alternative hypothesis[2] : means of at least two are not equal, is known as analysis of variance (ANOVA). The technique was introduced by R.A. Fisher.

In this chapter, we discuss two types of ANOVA

i) **One-way ANOVA:** Here only one factor or variable is considered for classification/analysis. This is always right-tailed test with the rejection region in the right-tail of the F-distribution curve.

ii) **Two-way ANOVA:** Here the data are classified on the basis of two factors.

7.2 Assumptions in ANOVA

The assumptions for ANOVA techniques are as follows:

1. **Independent samples:** The samples drawn from the populations are randomly selected and independent.
2. **Normal populations:** Samples are drawn from normally distributed populations.
3. **Equal variances:** The variances of the variable under consideration are the same for all the populations.

7.3 One-Way ANOVA

The data are classified/analyzed on the basis of only one factor or variable. For instance, in an agricultural firm, the quality of same kind of crops may be classified on the basis of different varieties of seeds or, on the basis of different varieties of fertilizers used but not both at a time (i.e., only one factor will be consider).

7.3.1 Methodology for One-Way ANOVA

Let there are k number of samples under consideration with sample size $n_i \ (i=1,2,....,k)$ and the following notations are used throughout the procedure. Thus n_i = Size of the sample i;

x_{ij} = The j^{th} value of the sample i

T_i = Total sum of the values of sample i $= \sum_{j=1}^{n_i} x_{ij}$

n = Number of the values in all samples = $n_1 + n_2 + ...$

T = Total sum of the values in all samples $= \sum_{i=1}^{k} T_i$

$\overline{x_i}$ = Mean of sample i $= \frac{T_i}{n_i}$

$\sum x^2$ = Sum of the squares of the values in all samples

The technique involves the following steps:

Step 1: Find mean of each sample $\overline{x_1}, \overline{x_2}, \cdots\cdots, \overline{x_k}$

Step 2: Find mean of the sample means (obtained in step-1) using

$$\overline{X} = \frac{\overline{x_1} + \overline{x_2} + \cdots\cdots + \overline{x_k}}{k}$$

Step 3: Calculate the sum of squares for variance between the samples (denoted as SSB), and mean square between samples (denoted by MSB) using

$$\text{SSB} = n_1\left(\overline{x_1} - \overline{X}\right)^2 + n_2\left(\overline{x_2} - \overline{X}\right)^2 + \cdots\cdots + n_k\left(\overline{x_k} - \overline{X}\right)^2$$

$$= \sum_{i=1}^{k} n_i\left(\overline{x_i} - \overline{X}\right)^2 = \sum_{i=1}^{k} n_i\,\overline{x}_i^{\,2} - n\,\overline{x}^2 = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

and $MSB = \frac{SSB}{k-1}$; (k–1) is the degrees of freedom (d.f.) between samples.

Step 4: Calculate the sum of squares for variance within samples(denoted as SSW), and mean square within samples (denoted as MSW) using

$$\text{SSW} = \sum_i \left(x_{1i} - \overline{x_1}\right)^2 + \sum_i \left(x_{2i} - \overline{x_2}\right)^2 + \cdots\cdots + \sum_i \left(x_{ki} - \overline{x_k}\right)^2 = \sum x^2 - \sum \frac{T_i^2}{n_i}$$

and $MSW = \frac{SSW}{n-k}$; n–k is the degrees of freedom within samples.

Step 5: Calculate total sum of squares (denoted by SST) using

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{X}\right)^2 = \sum_i \sum_j x_{ij}^2 - n\,\overline{X}^2 = \sum x^2 - \frac{T^2}{n};$$

From the above, it is clear that SST = SSB + SSW.

And the degrees of freedom for between and within sample

= (k–1) + (n–k) = (n–1), which is equal to the degrees of freedom for total variance.

This explains the additive property of the ANOVA technique.

Step 6: Now, the value of the test statistic F is given by the ratio MSB and MSW. Thus, the value of F statistic for an ANOVA test is given by

$$F_{Cal}=\frac{MSB}{MSW}$$

With degrees of freedom (k–1, n–k).

Step 7: Now, for a specific level of significance α, if calculated value of $F_{Cal}\geq F_{\alpha(k-1,n-k)}$, the Null hypothesis is rejected i.e., the population means are not equal.

And if $F_{Cal}<F_{\alpha(k-1,n-k)}$, the null hypothesis is accepted i.e., the population means are equal.

7.3.2 One-Way ANOVA Table

Source of Variation	**Sum of Squares (SS)**	**Degree of Freedom(d.f)**	**Mean Square (MS)**	**F-statistic (Calculated)**
Between Samples	$n_1\left(\overline{x_1}-\overline{X}\right)^2+n_2\left(\overline{x_2}-\overline{X}\right)^2+\cdots\cdots+n_k\left(\overline{x_k}-\overline{X}\right)^2$	k – 1	$\frac{SSB}{k-1}$	$\frac{\text{MSB}}{\text{MSW}}$
Within Samples	$\sum_i\left(x_{1i}-\overline{x_1}\right)^2+\sum_i\left(x_{2i}-\overline{x_2}\right)^2+\cdots\cdots+\sum_i\left(x_{ki}-\overline{x_k}\right)^2$ i = 1, 2, 3,.....	n – k	$\frac{SSW}{n-k}$	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{X}\right)^2$	n – 1		

7.3.3 Hypotheses for One-Way ANOVA and Conclusion

Null Hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ i.e., The means of all groups are equal.

Alternative Hypothesis H_1: At least two means are not equal.

Conclusion:

For the specific level of significance α, if calculated value of $F_{Cal} \geq F_{\alpha(k-1,n-k)}$ the Null hypothesis is rejected i.e., the population means are not equal.

And if $F_{Cal} < F_{\alpha(k-1,n-k)}$, the null hypothesis is accepted i.e., the population means are equal.

Example 1: Four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. As a result of dropouts from the experimental groups (due to sickness, transfer, etc.), the number of students varied from group to group. Do the data shown in Table below present sufficient evidence to indicate a difference in mean achievement for the four teaching techniques?

(Given that for 5% level of significance with df = (3,19), the value of $F_{0.05,\,(3,19)} = 3.13$)

Group 1	Group 2	Group 3	Group 4
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

i) State the null and alternative hypothesis to test the mean achievement is same for the four teaching techniques.

ii) Show the rejection region on the F-distribution for α = 0.05.

iii) Find SSB, SSW and SST.

iv) Find MSB and MSW.

v) Find the calculated value of F-statistic.

Solution:

i) The null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Against Alternative H_1 : All the four mean achievements is not equal.

ii) Here k = 4,

Sample size of the given four samples: $n_1=6$, $n_2=7$, $n_3=6$, $n_4=4$

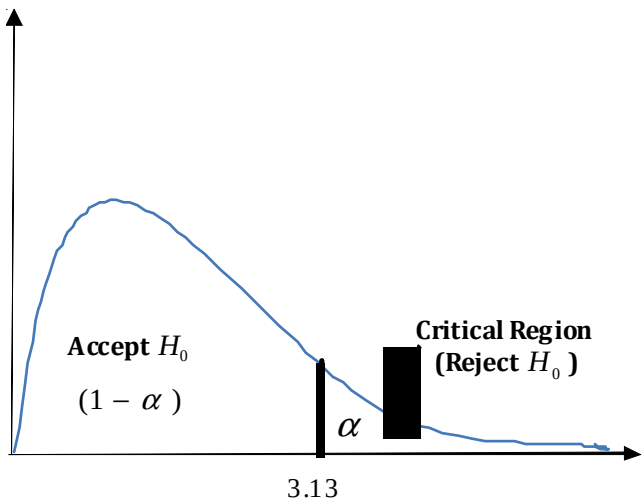
Total sample size $n=6+7+6+4=23$

Degree of freedom (d.f) for numerator is $k-1=3$

Degree of freedom (d.f) for denominator is $n-k=23-4=19$

Now given that, for $\alpha=0.05$ and df = (3,19), value of $F_{0.05,\,(3,19)}=3.13$.

Therefore, the rejection region lies to the right of F=3.13 in F distribution curve.



F- Distribution Curve

iii) Total of each group $T_1=454$, $T_2=549$, $T_3=425$, $T_4=351$

T = Total sum of all observation $= \sum_{i=1}^{4} T_i = 1779$

Mean of each sample $\overline{x}_1 = \frac{T_1}{n_1} = \frac{454}{6} = 75.67,$

$$\overline{x}_2 = \frac{T_2}{n_2} = \frac{549}{7} = 78.43,$$

$$\overline{x}_3 = \frac{T_3}{n_3} = \frac{425}{6} = 70.83,$$

$$\overline{x}_4 = \frac{T_4}{n_4} = \frac{351}{4} = 87.75$$

Mean of the sample means

$$\overline{X} = \frac{\overline{x}_1 + \overline{x}_2 + \overline{x}_3 + \overline{x}_4}{4} = \frac{75.67+78.43+70.83+87.75}{4} = 78.17$$

$$SSB = \sum_{i=1}^{4} \frac{T_i^2}{n_i} - \frac{T^2}{n} = \left[\frac{(454)^2}{6} + \frac{(549)^2}{7} + \frac{(425)^2}{6} + \frac{(351)^2}{4} \right] - \frac{(1779)^2}{23}$$

$$= 34352.67+43057.28+30104.17+30800.25-137601.78=712.59$$

$$SSW = \sum_i \left(x_{1i} - \overline{x}_1\right)^2 + \sum_i \left(x_{2i} - \overline{x}_2\right)^2 + \cdots\cdots + \sum_i \left(x_{ki} - \overline{x}_k\right)^2 = \sum x^2 - \sum \frac{T_i^2}{n_i}$$

$$=139511-138314.37=1196.63$$

$$SST = \sum x^2 - \frac{T^2}{n} = 139511 - 137601.78 = 1909.22$$

Hence, SSB+SSW=1909.22=SST

iv) $MSB=\frac{SSB}{k-1}=\frac{712.59}{3}=237.53$

$MSW=\frac{SSW}{n-k}=\frac{1196.63}{23-4}=62.98$

v) Now, the observed value of F-statistic for testing null hypothesis $H_0:\mu_1=\mu_2=\mu_3=\mu_4$ is given by $F_{Cal}=\frac{MSB}{MSW}=\frac{237.53}{62.98}=3.77$

At 5% level of significance with degrees of freedom $(k-1,n-k)=(3,19)$, the value of $F_{0.05,\,(3,19)}=3.13$.

Since, calculated value of $F_{Cal}=3.77\geq F_{0.05,(3,19)}=3.13$ at 5% level of significance and hence the Null hypothesis is rejected and concluded that there is sufficient evidence to indicate a difference in mean achievement among the four teaching procedures.

Example 2:

In a bookstall the sale of a book in a week, through three counters are recorded as given in the table.

Counter	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
I	9	10	8	7	11	12	
II	11	13	15	10	9	14	7
III	8	11	12	9	10	13	14

i) State the null and alternative hypothesis to test the mean sale.

ii) Show the rejection region on the F-distribution for $\alpha=0.05$.

iii) Find SSB, SSW and SST.

iv) Find MSB and MSW and also construct the ANOVA table.

v) Find the calculated value of F-statistic and Test the hypothesis that there is no difference between the mean sales of the book through each counter at 5% level of significance.

Solution:

Let μ_1, μ_2, μ_3 be the mean sale of the three counters respectively.

i) The null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ i.e., the mean sales are equal.

Against alternative hypothesis H_1 : At least two means are not equal.

ii) Here k = 3,

Sample size of the given three samples: $n_1=6$, $n_2=7$, $n_3=7$

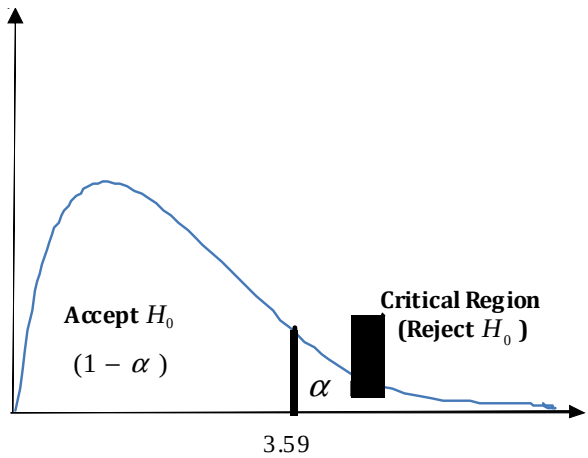
Total sample size $n=6+7+7=20$

Degree of freedom (d.f) for numerator is k – 1 = 2

Degree of freedom (d.f) for denominator is $n-k=20-3=17$

Now given that, for α = 0.05 and df = (2,17), value of $F_{0.05,\,(2,17)}=3.59$.

Therefore, the rejection region lies to the right of F=3.59 in F distribution curve.



F- Distribution Curve

iii) Total of each group $T_1=57$, $T_2=79$, $T_3=77$

T = Total sum of all observation $= \sum_{i=1}^{3} T_i = 213$

Mean of each sample $\bar{x}_1=\frac{T_1}{n_1}=\frac{57}{6}=9.50,$

$$\bar{x}_2=\frac{T_2}{n_2}=\frac{79}{7}=11.29,$$

$$\bar{x}_3=\frac{T_3}{n_3}=\frac{77}{7}=11.00,$$

Mean of the sample means $\bar{X}=\frac{\bar{x}_1+\bar{x}_2+\bar{x}_3}{3}=\frac{9.50+11.29+11}{3}=10.60$

$$SSB=\sum_{i=1}^{3}\frac{T_i^2}{n_i}-\frac{T^2}{n}=\left[\frac{(57)^2}{6}+\frac{(79)^2}{7}+\frac{(77)^2}{7}\right]-\frac{(213)^2}{20}$$

$$= 541.50 + 891.57 + 847.00 - 2268.45 =11.62$$

$$SSW=\sum_i\left(x_{1i}-\bar{x}_1\right)^2+\sum_i\left(x_{2i}-\bar{x}_2\right)^2+\cdots\cdots+\sum_i\left(x_{ki}-\bar{x}_k\right)^2=\sum x^2-\sum\frac{T_i^2}{n_i}$$

$$=2375.00-2280.07=94.93$$

$$SST=\sum x^2-\frac{T^2}{n}=2375.00-2268.45=106.55$$

Hence, SSB+SSW =106.55 = SST

iv) $MSB=\frac{SSB}{k-1}=\frac{11.62}{2}=5.81$

$$MSW=\frac{SSW}{n-k}=\frac{94.93}{17}=5.58$$

Source of Variation	Sum of Squares (SS)	Degree of Freedom (df)	Mean Square (MS)	F-Statistic (Calculated)
Between Samples	SSB = 11.62	2	$MSB=\frac{SSB}{k-1}=5.81$	$\frac{MSB}{MSW}=1.04$
Within Samples	SSW = 94.93	17	$MSW=\frac{SSW}{n-k}=5.58$	
Total	SST = 106.55	19		

v) Now, the observed value of F-statistic for testing null hypothesis $H_0:\mu_1=\mu_2=\mu_3$ is given by $F_{Cal}=\frac{MSB}{MSW}=\frac{5.81}{5.58}=1.04$

At 5% level of significance with degrees of freedom $(k-1,n-k)=(2,17)$, the value of $F_{0.05,(2,17)}=3.59$.

Since, calculated value of $F_{Cal}=1.04<F_{0.05,(2,17)}$ at level of significance and hence the Null hypothesis is accepted and concluded that there is no significant difference in mean sales among the three counters.

Example 3: Let μ_1,μ_2,μ_3 be respectively, the means of three normal distributions with a common but unknown variance. In order to test at 5% level of significance, the hypothesis $H_0:\mu_1=\mu_2=\mu_3$ against all possible alternative hypotheses, we take an independent random sample of size 4 from each of these distributions. Determine whether we accept or reject H_0 if the observed values from these three distributions are respectively,

X_1 :	5	9	6	8
X_2 :	11	13	10	12
X_3 :	10	6	9	9

Solution:

i) The null hypothesis $H_0:\mu_1=\mu_2=\mu_3$

Against Alternative H_1 : At least two of μ_1, μ_2, μ_3 are not equal.

ii) Here k = 3,

Sample size of the given four samples: $n_1=4, n_2=4, n_3=4$

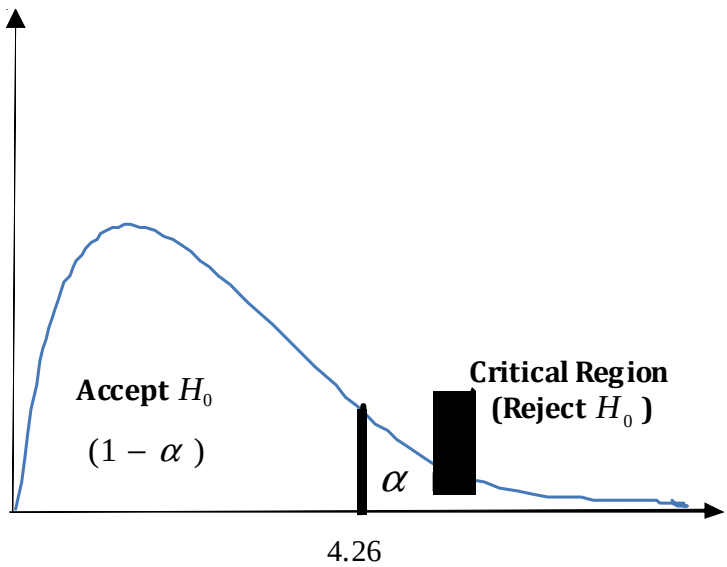
Total sample size $n=4+4+4=12$

Degree of freedom (d.f) for numerator is k – 1 = 2

Degree of freedom (d.f) for denominator is $n-k=12-3=9$

Now given that, for α = 0.05 and df = (2,9), value of $F_{0.05,(2,9)}=4.26$.

Therefore, the rejection region lies to the right of F=4.26 in F distribution curve.



F- Distribution Curve

iii) Total of each group $T_1=28, T_2=46, T_3=34$

T = Total sum of all observation $= \sum_{i=1}^{3} T_i = 108$

Mean of each sample $\bar{x}_1=\frac{T_1}{n_1}=\frac{28}{4}=7,$

$$\overline{x_2}=\frac{T_2}{n_2}=\frac{46}{4}=11.50,$$

$$\overline{x_3}=\frac{T_3}{n_3}=\frac{34}{4}=8.50$$

Mean of the sample means $\overline{X}=\frac{\overline{x_1}+\overline{x_2}+\overline{x_3}}{3}=\frac{7+11.50+8.50}{3}=9$

$$SSB=\sum_{i=1}^{3}\frac{T_i^2}{n_i}-\frac{T^2}{n}=\left[\frac{(28)^2}{4}+\frac{(46)^2}{4}+\frac{(34)^2}{4}\right]-\frac{(108)^2}{12}$$

$$= 196 + 529 +289 -972 = 42$$

$$SSW=\sum_i\left(x_{1i}-\overline{x_1}\right)^2+\sum_i\left(x_{2i}-\overline{x_2}\right)^2+\cdots\cdots+\sum_i\left(x_{ki}-\overline{x_k}\right)^2=\sum x^2-\sum\frac{T_i^2}{n_i}$$

$$= 1038 - 1014 = 24$$

$$SST=\sum x^2-\frac{T^2}{n}=1038-972=66$$

Hence, SSB+SSW= 66 =SST

Source of Variation	**Sum of Squares (SS)**	**Degree of Freedom (d.f)**	**Mean Square (MS)**	**F-statistic (Calculated)**
Between Samples	$SSB=\sum_{i=1}^{3}\frac{T_i^2}{n_i}-\frac{T^2}{n}$ = 42	k–1=2	$MSB=\frac{SSB}{k-1}=\frac{42}{2}=21$	$\frac{MSB}{MSW}=\frac{21}{2.67}=7.87$
Within Samples	$SSW=\sum x^2-\sum\frac{T_i^2}{n_i}$ = 24	n–k=9	$MSW=\frac{SSW}{n-k}=\frac{24}{9}=2.67$	
Total	$SST=\sum x^2-\frac{T^2}{n}$ = 66	n–1=11		

At 5% level of significance with degrees of freedom $(k-1, n-k)=(2,9)$, the tabulated value of $F_{0.05,(2,9)}=4.26$.

Since, calculated value of $F_{Cal}=7.87>F_{0.05,(2,9)}=4.26$ at 5% level of significance and hence the Null hypothesis is rejected.

Example 4: With the ongoing energy crisis, researchers for the major oil companies are attempting to find alternative sources of oil. It is known that some types of shale contain small amounts of oil that feasibly (if not economically) could be extracted. Four methods have been developed for extracting oil from shale, and the government has decided that some experimentation should be done to determine whether the methods differ significantly in the average amount of oil that each can extract from the shale. Method 4 is known to be the most expensive method to implement, and method 1 is the least expensive, so inferences about the differences in performance of these two methods are of particular interest. Sixteen bits of shale (of the same size) were randomly subjected to the four methods, with the results shown in the accompanying table (the units are in liters per cubic meter). All inferences are to be made with α = 0.05.

Method 1	Method 2	Method 3	Method 4
3	2	5	5
2	2	2	2
1	4	5	4
2	4	1	5

Assuming that the 16 experimental units were as alike as possible, implement the appropriate ANOVA to determine whether there is any significant difference among the mean amounts extracted by the four methods. Use α =0.05.

Solution:

i) The null hypothesis $H_0: \mu_1=\mu_2=\mu_3=\mu_4$

Against Alternative H_1 : At least two of $\mu_1, \mu_2, \mu_3, \mu_4$ are not equal.

ii) Here k = 4,

Sample size of the given four samples: $n_1=4,\ n_2=4,\ n_3=4, n_4=4$

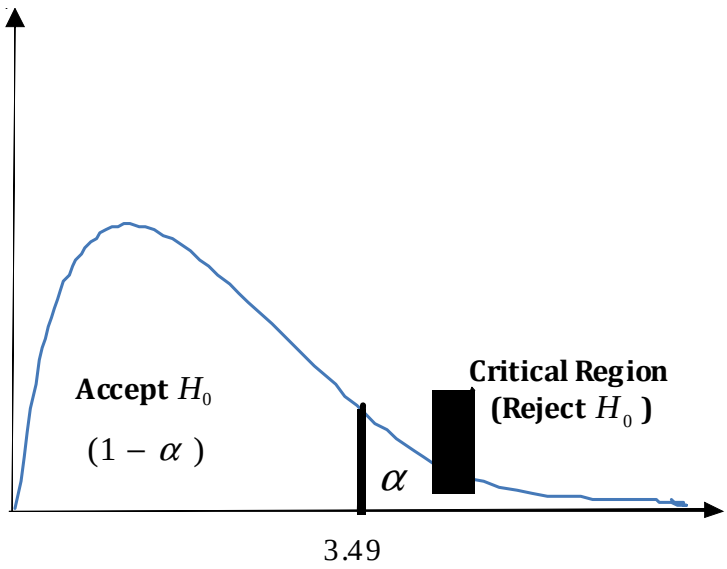
Total sample size $n=4+4+4+4=16$

Degree of freedom (d.f) for numerator is k – 1 = 3

Degree of freedom (d.f) for denominator is $n-k=16-4=12$

Now given that, for α = 0.05 and df = (3,12), value of $F_{0.05,\,(3,12)}=3.49$.

Therefore, the rejection region lies to the right of F= 3.49 in F distribution curve.



F- Distribution Curve

(iii) Total of each group $T_1=8,\ T_2=12,\ T_3=13, T_4=16$

T = Total sum of all observation $= \sum_{i=1}^{4} T_i = 49$

Now, $\sum_{i=1}^{4} \frac{T_i^2}{n_i} = \left[\frac{8^2}{4}+\frac{12^2}{4}+\frac{13^2}{4}+\frac{16^2}{4}\right] = 16+36+42.25+64 = 158.25$

$\frac{T^2}{n} = \frac{49^2}{16} = 150.06$

Hence, $SSB=\sum_{i=1}^{4}\frac{T_i^2}{n_i}-\frac{T^2}{n}=\left[\frac{8^2}{4}+\frac{12^2}{4}+\frac{13^2}{4}+\frac{16^2}{4}\right]-\frac{49^2}{16}$

$= 16 + 36 + 42.25 + 64 - 150.06 = 8.19$

$SSW=\sum x^2-\sum\frac{T_i^2}{n_i}=183-158.25=24.75$

$SST=\sum x^2-\frac{T^2}{n}=183-150.06=32.94$

Hence, SSB + SSW = 32.94 = SST

Source of Variation	**Sum of Squares (SS)**	**Degree of Freedom (d.f)**	**Mean Square (MS)**	**F-statistic (Calculated)**
Between Samples	$SSB=\sum_{i=1}^{3}\frac{T_i^2}{n_i}-\frac{T^2}{n}$ $= 8.19$	k–1=3	$MSB=\frac{SSB}{k-1}=\frac{8.19}{3}=2.73$	$\frac{MSB}{MSW}=\frac{2.73}{2.06}=1.33$
Within Samples	$SSW=\sum x^2-\sum\frac{T_i^2}{n_i}$ $= 24.75$	n–k=12	$MSW=\frac{SSW}{n-k}=\frac{24.75}{12}=2.06$	
Total	$SST=\sum x^2-\frac{T^2}{n}$ $= 32.94$	n–1=15		

At 5% level of significance with degrees of freedom $(k-1,n-k)=(3,12)$, the tabulated value of $F_{0.05,\,(3,12)}=3.49$.

Since, calculated value $F_{Cal}=1.33<F_{0.05,(3,12)}=3.49$ at 5% level of significance hence the Null hypothesis is accepted and concluded that there is no significant difference among the mean amounts extracted by the four methods.

Example 5:

Under normal conditions, is the average body temperature the same for men and women?

Medical researchers interested in this question collected data from a large number of men and women, and random samples from that data are presented in the accompanying table. Is there sufficient evidence to indicate that mean body temperatures differ for men and women?

Body Temperatures (°F)	
Men	**Women**
96.9	97.8
97.4	98.0
97.5	98.2
97.8	98.2
97.8	98.2
97.9	98.6
98.0	98.8
98.6	99.2
98.8	99.4

Solution:

i) The null hypothesis $H_0 : \mu_1=\mu_2$

Against Alternative H_1 : At least two of μ_1, μ_2 are not equal.

ii) Here, k = 2

Sample size of the given two samples: $n_1=9, n_2=9$

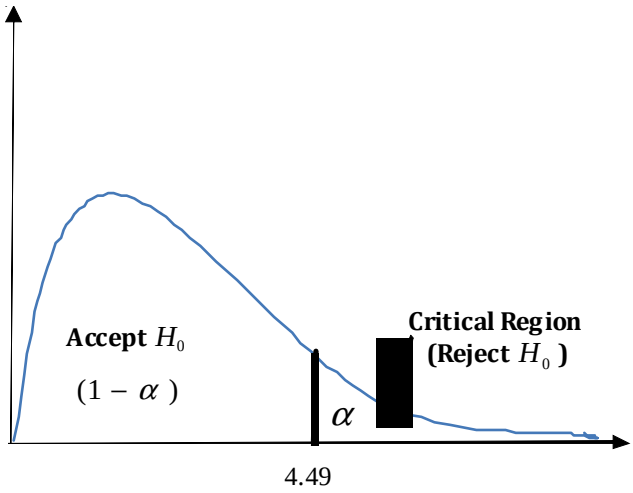
Total sample size $n=9+9=18$

Degree of freedom (d.f) for numerator is k – 1 = 1

Degree of freedom (d.f) for denominator is $n-k=18-2=16$

Now given that, for α = 0.05 and df = (1,16), value of $F_{0.05,\,(1,16)}=4.49$.

Therefore, the rejection region lies to the right of F = 4.49 in F distribution curve.



F- Distribution Curve

(iii) Total of each group $T_1=880.70$, $T_2=886.40$

T = Total sum of all observation $=\sum_{i=1}^{2} T_i = 1767.10$

Now, $\sum_{i=1}^{2}\frac{T_i^2}{n_i}=\left[\frac{(880.70)^2}{9}+\frac{(886.40)^2}{9}\right]=86181.39+87300.55=173481.94$

$$\frac{T^2}{n}=\frac{(1767.10)^2}{18}=173480.13$$

Hence, $$SSB=\sum_{i=1}^{2}\frac{T_i^2}{n_i}-\frac{T^2}{n}=173481.94-173480.13=1.81$$

$$SSW=\sum x^2-\sum\frac{T_i^2}{n_i}=173487.07-173481.94=5.13$$

$$SST=\sum x^2-\frac{T^2}{n}=173487.07-173480.13=6.94$$

Hence, SSB + SSW = 6.94 = SST

Source of Variation	**Sum of Squares (SS)**	**Degree of Freedom (d.f)**	**Mean Square (MS)**	**F-statistic (Calculated)**
Between Samples	$SSB=\sum_{i=1}^{3}\frac{T_i^2}{n_i}-\frac{T^2}{n}$ = 1.81	k–1=1	$MSB=\frac{SSB}{k-1}=1.81$	$\frac{MSB}{MSW}=\frac{1.81}{0.32}$ = 5.66
Within Samples	$SSW=\sum x^2-\sum\frac{T_i^2}{n_i}$ = 5.13	n–k=16	$MSW=\frac{SSW}{n-k}=\frac{5.13}{16}$ = 0.32	
Total	$SST=\sum x^2-\frac{T^2}{n}$ = 6.94	n–1=17		

At 5% level of significance with degrees of freedom $(k-1,n-k)=(1,16)$, the tabulated value of $F_{0.05,\,(1,16)}=4.49$.

Since, calculated value $F_{Cal}=5.56>F_{0.05,(1,16)}=4.49$ at 5% level of significance hence the Null hypothesis is rejected.

Example 6:

A dealer has in stock three cars (models A, B, and C) of the same make but different models. Wishing to compare mileage obtained for these different models, a

customer arranged to test each car with each of three brands of gasoline (brands X, Y, and Z). In each trial, a gallon of gasoline was added to an empty tank, and the car was driven without stopping until it ran out of gasoline. The accompanying table shows the number of miles covered in each of the nine trials.

	Distance (miles)		
Brand of Gasoline	**Model A**	**Model B**	**Model C**
X	22.4	17.0	19.2
Y	20.8	19.4	20.2
Z	21.5	18.7	21.2

Should the customer conclude that the different car models differ in mean gas mileage? Test at the α = .05 level.

Solution:

i) The null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$

Against Alternative H_1. At least two of μ_1, μ_2, μ_3 are not equal.

ii) Here, k = 3

Sample size of the given three samples: $n_1 = 3, n_2 = 3, n_3 = 3$

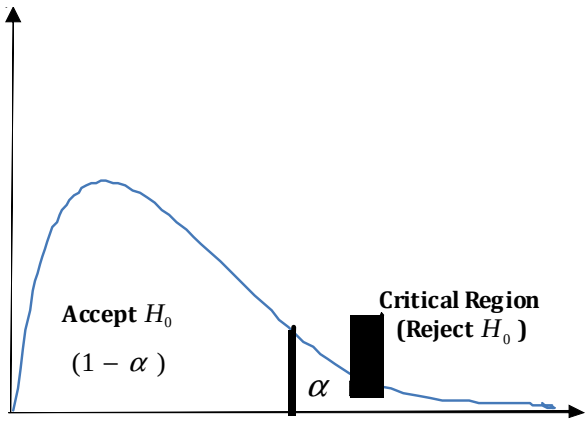
Total sample size $n = 3+3+3 = 9$

Degree of freedom (d.f) for numerator is k – 1 = 2

Degree of freedom (d.f) for denominator is $n - k = 9 - 3 = 6$

Now given that, for α = 0.05 and df = (2,6), value of $F_{0.05,(2,6)} = 5.14$.

Therefore, the rejection region lies to the right of F = 5.14 in F distribution curve.



5.14

F- Distribution Curve

(iii) Total of each group T_1=64.70, T_2=55.10, T_3=60.60

T = Total sum of all observation $= \sum_{i=1}^{3} T_i$ =180.40

Now, $$\sum_{i=1}^{3} \frac{T_i^2}{n_i} = \left[\frac{(64.70)^2}{3} + \frac{(55.10)^2}{3} + \frac{(60.60)^2}{3} \right] = 3631.49$$

$$\frac{T^2}{n} = \frac{(180.40)^2}{9} = 3616.02$$

Hence, $$SSB = \sum_{i=1}^{3} \frac{T_i^2}{n_i} - \frac{T^2}{n} = 3631.49 - 3616.02 = 15.47$$

$$SSW = \sum x^2 - \sum \frac{T_i^2}{n_i} = 3637.82 - 3631.49 = 6.33$$

$$SST = \sum x^2 - \frac{T^2}{n} = 3637.82 - 3616.02 = 21.80$$

Hence, SSB + SSW = 21.80 = SST

Source of Variation	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Square (MS)	F-statistic (Calculated)
Between Samples	$SSB=\sum_{i=1}^{3}\frac{T_i^2}{n_i}-\frac{T^2}{n}$ = 15.47	k–1=2	$MSB=\frac{SSB}{k-1}=\frac{15.47}{2}=7.74$	$\frac{MSB}{MSW}=\frac{7.74}{1.06}$ = 7.30
Within Samples	$SSW=\sum x^2-\sum\frac{T_i^2}{n_i}$ = 6.33	n–k=6	$MSW=\frac{SSW}{n-k}=\frac{6.33}{6}=1.06$	
Total	$SST=\sum x^2-\frac{T^2}{n}$ = 21.80	n–1=8		

At 5% level of significance with degrees of freedom $(k-1, n-k)=(2,6)$, the tabulated value of $F_{0.05,(2,6)}=5.14$.

Since, calculated value $F_{Cal}=7.30>F_{0.05,(2,6)}=5.14$ at 5% level of significance hence the Null hypothesis is rejected.

Example 7:

In the hope of attracting more riders, a city transit company plans to have express bus service from a suburban terminal to the downtown business district. These buses should save travel time. The city decides to perform a study of the effect of four different plans (such as a special bus lane and traffic signal progression) on the travel time for the buses. Travel times (in minutes) are measured for several weekdays during a morning rush-hour trip while each plan is in effect. The results are recorded in the following table.

Plan			
1	**2**	**3**	**4**
27	25	34	30
25	28	29	33
29	30	32	31
26	27	31	
	24	36	

Is there evidence of a difference in the mean travel times for the four plans? Use $\alpha = 0.01$.

Solution:

i) The null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Against Alternative H_1 : At least two of $\mu_1, \mu_2, \mu_3, \mu_4$ are not equal.

ii) Here, k = 4

Sample size of the given four samples: $n_1 = 4,\ n_2 = 5, n_3 = 5, n_4 = 3$

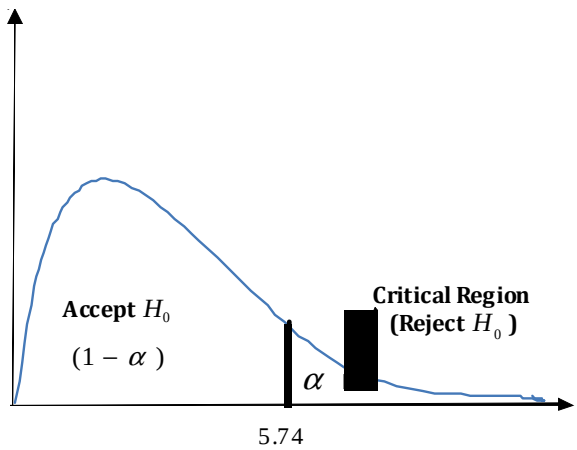
Total sample size $n = 4+5+5+3 = 17$

Degree of freedom (d.f) for numerator is k – 1 = 3

Degree of freedom (d.f) for denominator is $n - k = 17 - 4 = 13$

Now given that, for $\alpha = 0.01$ and df = (3,13), value of $F_{0.01,\,(3,13)} = 5.74$.

Therefore, the rejection region lies to the right of F = 5.74 in F distribution curve.



F- Distribution Curve

(iii) Total of each group $T_1=107,\ T_2=134, T_3=162, T_4=94$

T = Total sum of all observation $= \sum_{i=1}^{4} T_i = 497$

Now, $\sum_{i=1}^{4} \frac{T_i^2}{n_i} = \left[\frac{(107)^2}{4} + \frac{(134)^2}{5} + \frac{(162)^2}{5} + \frac{94^2}{3}\right] = 14647.58$

$$\frac{T^2}{n} = \frac{(497)^2}{17} = 14529.94$$

Hence, $$SSB = \sum_{i=1}^{4} \frac{T_i^2}{n_i} - \frac{T^2}{n} = 14647.58 - 14529.94 = 117.64$$

$$SSW = \sum x^2 - \sum \frac{T_i^2}{n_i} = 14713 - 14647.58 = 65.42$$

$$SST = \sum x^2 - \frac{T^2}{n} = 14713 - 14529.94 = 183.06$$

Hence, SSB + SSW = 183.06 = SST

Source of Variation	**Sum of Squares (SS)**	**Degree of Freedom (d.f)**	**Mean Square (MS)**	**F-statistic (Calculated)**
Between Samples	$SSB = \sum_{i=1}^{3} \frac{T_i^2}{n_i} - \frac{T^2}{n}$ = 117.64	k–1=3	$MSB = \frac{SSB}{k-1} = \frac{117.64}{3}$ =39.21	$\frac{MSB}{MSW} = \frac{39.21}{5.03}$ = 7.80
Within Samples	$SSW = \sum x^2 - \sum \frac{T_i^2}{n_i}$ = 65.42	n–k=13	$MSW = \frac{SSW}{n-k} = \frac{65.42}{13}$ = 5.03	
Total	$SST = \sum x^2 - \frac{T^2}{n}$ = 183.06	n–1=16		

At 1% level of significance with degrees of freedom $(k-1, n-k)=(3,13)$, the tabulated value of $F_{0.01,\,(3,13)}=5.74$.

Since, calculated value $F_{Cal}=7.80>F_{0.01,(3,13)}=5.74$ at 1% level of significance hence the Null hypothesis is rejected.

7.4 Two-Way ANOVA

In case of two-way ANOVA technique, the observations are classified on the basis of two factors. For example, the sale of a product depends on several factors such as efficiency of salesman, location of shop, quality of the product etc and thus the sale of a particular product may be classified on the basis of efficiency of salesman and also on the basis of location of shop. In an agricultural firm, products may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used.

In this section, we discuss the two-way ANOVA technique with one observation per cell only.

7.4.1 Methodology for Two-Way ANOVA

Let us consider model with p groups of factor A and q groups of factor B as given below.

Factor(s)	B_1	B_2	$\cdots$	B_q	Total ($x_{i\bullet}$)
A_1	x_{11}	x_{12}	$\cdots$	x_{1q}	$x_{1\bullet}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2q}	$x_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_p	x_{p1}	x_{p2}	$\cdots$	x_{pq}	$x_{p\bullet}$
Total ($x_{\bullet j}$)	$x_{\bullet 1}$	$x_{\bullet 2}$		$x_{\bullet q}$	$x_{\bullet\bullet}$

Where x_{ij} is the observation of $(i,j)^{th}$ cell, which belongs to i^{th} class of factor A and class j^{th} of factor B.

The model is $x_{ij}=\mu+\alpha_i+\beta_j+e_{ij}$.

Where μ is general mean effect,

α_i is the effect of i^{th} class of factor A,

β_j is the effect of j^{th} class of factor B,

e_{ij} is the error component and it is independent and normally distributed with mean 0 (zero) and variance σ^2.

$n = pq$ is the total observations.

Total Sum of Squares = Sum of Squares due to factor A + Sum of Squares due to factor B + Sum of Squares due to Error

i.e., $SST=SSA+SSB+SSE$

Thus, $d.f(SST)=d.f(SSA)+d.f(SSB)+d.f(SSE)$

i.e., $n-1=(p-1)+(q-1)+(p-1)(q-1)$

where, Total of all observation = $T=\sum_{i=1}^{p}\sum_{j=1}^{q}x_{ij}=x_{\bullet\bullet}$

$$SST=\sum_{i=1}^{p}\sum_{j=1}^{q}x_{ij}^{2}-\frac{T^2}{n}$$

$$SSA=\frac{1}{q}\sum_{i=1}^{p}x_{i\bullet}^{2}-\frac{T^2}{n}$$

$$SSB=\frac{1}{p}\sum_{j=1}^{q}x_{\bullet j}^{2}-\frac{T^2}{n}$$

$$SSE=SST-SSA-SSB$$

$x_{i\bullet}=Total\ of\ i^{th}\ level\ of\ \ factor\ A$

$x_{\bullet j}=Total\ of\ \ j^{th}\ level\ of\ \ factor\ B$

7.4.2 Two-Way ANOVA Table

Source of Variation	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Square (MS)	F-ratio
Between Columns/ Factor A	$SSA = \frac{1}{q}\sum_{i=1}^{p} x_{i\bullet}^2 - \frac{T^2}{n}$	$p-1$	$MSA = \frac{SSA}{p-1}$	$F_A = \frac{MSA}{MSE}$
Between Rows/ Factor B	$SSB = \frac{1}{p}\sum_{j=1}^{q} x_{\bullet j}^2 - \frac{T^2}{n}$	$q-1$	$MSB = \frac{SSB}{q-1}$	$F_B = \frac{MSB}{MSE}$
Residual/ Error	$SST - SSA - SSB$	$(p-1)(q-1)$	$MSE = \frac{SSE}{(p-1)(q-1)}$	
Total	$SST = \sum_{i=1}^{p}\sum_{j=1}^{q} x_{ij}^2 - \frac{T^2}{n}$	$n-1$		

Thus, $F_A = \frac{MSA}{MSE} \sim F_{(p-1),(p-1)(q-1)}$;

$F_B = \frac{MSB}{MSE} \sim F_{(q-1),(p-1)(q-1)}$

7.4.3 Hypotheses for Two-Way ANOVA and Conclusion

Null Hypothesis:

$H_{0A}: \mu_{1A} = \mu_{2A} = \cdots = \mu_{pA}$ i.e., the mean of all groups of factor A is same.

$H_{0B}: \mu_{1B} = \mu_{2B} = \cdots = \mu_{qB}$ i.e., the mean of all groups of factor B is same.

Alternative Hypothesis:

H_{1A} : Mean of at least two groups of factor A are not same.

H_{1B} : Mean of at least two groups of factor B are not same.

Conclusion:

The F-ratio is used to compare whether the difference among several sample means is significant or is just due to random causes/chance.

If $F_A > F_{(p-1),(p-1)(q-1)}\alpha$, then H_{0A} is rejected at 100α% level of significance and we conclude that groups differ significantly, otherwise H_{0A} accepted.

If $F_B > F_{(q-1),(p-1)(q-1)}\alpha$, then H_{0B} is rejected at 100α% level of significance and we conclude that groups differ significantly, otherwise H_{0B} accepted.

Example 1:

Three testing agencies determine the alcohol contents of samples of sanitizer. For this purpose each agency has taken 4 packets. The results are given below:

	Packet			
Agency	**I**	**II**	**III**	**IV**
A	9	10	9	10
B	12	11	9	11
C	11	12	10	12

i) State the null and alternative hypothesis.

ii) Compute SST, SSA, SSB and SSE.

iii) Construct the ANOVA table and find the calculated values of F-ratio.

iv) Test whether there is any significant difference among packets and among agencies at 5% significance of level.

Solution:

i) $H_{0A}: \mu_{1A} = \mu_{2A} = \mu_{3A}$ i.e., there is no difference among the means of agencies.

Against H_{1A}: at least two $\mu_{iA}\,(i = 1, 2, 3)$ are not same.

$H_{0B}: \mu_{1B}=\mu_{2B}=\mu_{3B}=\mu_{4B}$ i.e., there is no difference among the means of packets.

Against H_{1B}: At least two $\mu_{jB}\ (j=1,2,3,4)$ are not same.

ii)

	Packet				
Agency	**I**	**II**	**III**	**IV**	**Total ($x_{i\bullet}$)**
A	9	10	9	10	$x_{1\bullet}=38$
B	12	11	9	11	$x_{2\bullet}=43$
C	11	12	10	12	$x_{3\bullet}=45$
Total ($x_{\bullet j}$)	$x_{\bullet 1}=32$	$x_{\bullet 2}=33$	$x_{\bullet 3}=28$	$x_{\bullet 4}=33$	$x_{\bullet\bullet}=126$

Here, $p=3, q=4, n=pq=12$

Therefore, $SST=\sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij}^{\ 2}-\frac{T^2}{n}=[81+100+81+...+144]-\frac{126^2}{12}=1338-1323=15$

$$SSA=\frac{1}{4}\sum_{i=1}^{3} x_{i\bullet}^2-\frac{T^2}{n}=\frac{1}{4}\left[38^2+43^2+45^2\right]-\frac{126^2}{12}=1329.5-1323=6.5$$

$$SSB=\frac{1}{3}\sum_{j=1}^{4} x_{\bullet j}^2-\frac{T^2}{n}=\frac{1}{3}\left[32^2+33^2+28^2+33^2\right]-\frac{126^2}{12}=1328.67-1323=5.67$$

$SSE=SST-SSA-SSB=15-6.5-5.67=2.83$

iii) ANOVA Table:

Source of Variation	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Square (MS)	F-ratio	Tabulated F(0.05)
Between Columns/ Factor A	6.5	2	3.25	6.91	5.14
Between Rows/ Factor B	5.67	3	1.89	4.02	4.76
Residual/ Error	2.83	6	0.47		
Total	15.0	11			

iv) Now, the tabulated values are $F_{2,6}(0.05)=5.14$ *and* $F_{3,6}(0.05)=4.76$

The F-ratio = 6.91 (>5.14) lies in the critical region at 5% level of significance, therefore, we conclude that the mean alcohol content as determined by 3 testing agencies are not equal.

The F-ratio= 4.02 (<4.76) does not lie in the critical region at 5% level of significance, therefore, we conclude that the alcohol content of the 4 packets may not different from one another.

Example 2:

Three varieties of seeds are used and the production (in metric tons) per acre of land using four types of fertilizers are given in the following table.

	Seeds		
Fertilizers	**I**	**II**	**III**
A	6	5	5
B	7	5	4
C	3	3	3
D	8	7	4

i) State the null and alternative hypothesis.

ii) Compute SST, SSA, SSB and SSE.

iii) Construct the ANOVA table and find the calculated values of F-ratio.

iv) Test whether there is any significant difference among fertilizers and seeds at 5% significance of level.

Solution:

i) $H_{0A}: \mu_{1A}=\mu_{2A}=\mu_{3A}=\mu_{4A}$ i.e., there is no difference among the means of fertilizers.

Against H_{1A} : at least two $\mu_{iA}\,(i=1,2,3,4)$ are not same.

$H_{0B}: \mu_{1B}=\mu_{2B}=\mu_{3B}$ i.e., there is no difference among the means of Seeds.

Against H_{1B} : At least two $\mu_{jB}\,(j=1,2,3)$ are not same.

ii)

	Seed			
Fertilizer	**I**	**II**	**III**	**Total ($x_{i\bullet}$)**
A	6	5	5	$x_{1\bullet}=16$
B	7	5	4	$x_{2\bullet}=16$
C	3	3	3	$x_{3\bullet}=9$
D	8	7	4	$x_{4\bullet}=19$
Total ($x_{\bullet j}$)	$x_{\bullet 1}=24$	$x_{\bullet 2}=20$	$x_{\bullet 3}=16$	$x_{\bullet\bullet}=60$

Here, $p=4, q=3, n=pq=12$

Therefore, $SST=\sum_{i=1}^{4}\sum_{j=1}^{3} x_{ij}^{\,2}-\frac{T^2}{n}=[36+25+25+...+16]-\frac{60^2}{12}=332-300=32$

$$SSA=\frac{1}{3}\sum_{i=1}^{4} x_{i\bullet}^2-\frac{T^2}{n}=\frac{1}{3}\left[16^2+16^2+9^2+19^2\right]-\frac{60^2}{12}=318-300=18$$

$$SSB=\frac{1}{4}\sum_{j=1}^{3} x_{\bullet j}^2-\frac{T^2}{n}=\frac{1}{4}\left[24^2+20^2+16^2\right]-\frac{60^2}{12}=308-300=8$$

$SSE = SST - SSA - SSB = 32 - 18 - 8 = 6$

iii) ANOVA Table:

Source of Variation	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Square (MS)	F-ratio	Tabulated F(0.05)
Between Columns/ Factor	18	3	6	6	4.76
Between Rows/ Factor	8	2	4	4	5.14
Residual/ Error	6	6	1		
Total	32	11			

iv) Now, the tabulated values are $F_{3,6}(0.05)=4.76$ and $F_{2,6}(0.05)=5.14$

The F-ratio = 6 (>4.76) lies in the critical region at 5% level of significance, therefore, we conclude that the variety of fertilizers is significant.

The F-ratio= 4 (<5.14) does not lie in the critical region at 5% level of significance; therefore, we conclude that the differences of seeds is insignificant.

7.5 Summary

In this unit, we have presented the one-way ANOVA and two-way ANOVA with only one observation per cell. As the number of populations involved is more than two, F-distribution is used for testing the equality of means of all populations comparing sample variances. We have discussed how to test null hypothesis (all the population means are equal) against the alternative hypothesis (means of at least of two are not equal) and also explained with number of examples.

7.5 Self-Assessment Questions

1. A random sample of five motor-car tyres is taken from each of the three brands manufactured by three companies. The lifetime of these tyres is shown below. On the basis of the data, test whether the average lifetime of the 3 brands of tyres are equal or not at 1% level of significance.

Brand A	Brand B	Brand C
35	32	34
34	32	33
34	31	32
33	28	32
34	29	33

2. The lifetime (in Hrs.) of four brands of batteries are given in the table.

Brand A	Brand B	Brand C	Brand D
29	39	27	25
40	36	33	28
35	42	28	32
27	28	34	38
32	31	36	27

Test whether the mean lifetime of four brands of batteries are equal or not at 5% level of significance.

3. Find the missing entries of the one-way ANOVA table given below.

Source of Variation	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Square (MS)	F-statistic
Between	117.64	****	39.21	7.80
Within	*****	13	*****	
Total	*****	*****		

4. Complete the one-way ANOVA table given below.

Source of Variation	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Square (MS)	F-statistic (Calculated)
Between	102.4	****	51.2	10.6
Within	****	12	****	
Total	****	****		

5. Complete the two-way ANOVA table given below.

Source of Variation	Sum of Squares (SS)	Degree of Freedom (d.f)	Mean Square (MS)	F-ratio
Between Columns/ Factor A	6.5	***	***	6.91
Between Rows/ Factor B	***	3	***	4.02
Residual/ Error	2.83	***	0.47	
Total	15.0	11		

6. The temperatures, in Celsius, at three locations in the engine of a vehicle are measured after each of four test runs and the data are as follows. Making the usual assumptions for a two-way analysis of variance, test the hypothesis that there is no systematic difference in temperatures between the three locations. Use the 5% level of significance.

Location	Run 1	Run 2	Run 3	Run 4
A	72.8	77.3	82.9	69.4
B	71.5	72.4	80.7	67.0
C	70.8	74.0	79.1	69.0

References:

Dennis D. Wackerly, William Mendenhall III, Richard L. Scheaffer. Mathematical Statistics with Applications, Seventh Edition, Thomson Brooks/Cole.

Robert V. Hogg, Joseph W. McKean, Allen T. Craig. Introduction to Mathematical Statistics, Eighth Edition, Pearson.

Unit 8 ❑ Statistical Quality Control

Structure

8.0 Objectives

Quality and productivity have become essential for organizations to be competitive in today's global economy. As a result, an increased emphasis falls on methods for maintaining and monitoring quality standards. After studying the present unit, you will be able to (i) understand the concept of statistical quality control, (ii) explain the importance of statistical quality control and (iii) discuss the different product control and process control techniques.

8.1 Introduction

In the present day situation where there is a highly competitive market, it has become absolutely necessary for every manufacturer to maintain a close supervision on the quality of articles produced. If the customers are not satisfied with the quality of the products, it will be very difficult for the producers to remain in the competition. Statistical quality control provides certain statistical techniques used for maintaining conformity with certain standards in a continuous flow of manufactured products. Thus, the application of statistical quality control is inevitable in every manufacturing organization. In fact, at present it has also become an integral part of management control system.

8.2 Definition of Statistical Quality Control

Statistical Quality Control (SQC) may be defined as the statistical techniques employed for determining the extent to which quality goals are being met without necessarily checking every item produced, for indicating whether or not the variations which occur are exceeding normal expectations and for taking decisions regarding acceptance or rejection of a particular product. In the words of Alford and Beatly, SQC may be defined as that industrial management technique by means of which products of uniform acceptable quality are manufactured. It is mainly concerned with making things right rather than discovering and rejecting those made wrong.

8.3 Need For Statistical Quality Control

(i) SQC assures the customers to rely on the quality standard of the products consumed by them.

(ii) It makes the employees of the firm quality conscious. As a result, the employees

pay due attention to the maintenance of quality standard of product manufactured in the factory.

(iii) It protects the manufacturers as well as the customers against heavy losses due to rejection of large quantity of products.

(iv) It acts as a guide for setting up of a future standard of quality when the existing control limits are required to be amended.

(v) It helps in establishing goodwill of the products.

8.4 Causes of Variation

It is desirable in a production unit that all the products should be produced according to the prescribed specification. But in practice, it is very rare that all the products produced in the production unit are of exactly same quality. Some products may be slightly below the prescribed standard and some may be slightly above the same. It leads to a search in the possible causes of variation in the products. Generally, the variation in the quality of products arises due to two distinct types of causes - (i) Chance causes and (ii) Assignable causes.

Let us discuss these two one by one.

8.4.1 Chance Causes

Some small variations in the quality of products in a manufacturing process are inherent and cannot be prevented altogether in any way. This type of variation is generated by several independent factors which are known as 'chance causes'. If the variations in the quality of products in a production process are solely due to chance causes, the process is said to be 'under control' or 'in a state of statistical control'.

8.4.2 Assignable Casuses

The major variations in the quality of products are generally caused by defects and faults in the production design and manufacturing process. The factors responsible for such variations in the quality of products are known as 'assignable causes'. These causes are not inherent and can be prevented. The value of quality control lies in the fact that variations arising out of assignable causes can be quickly detected; even probable variations due to these causes may often be identified before the products become defective.

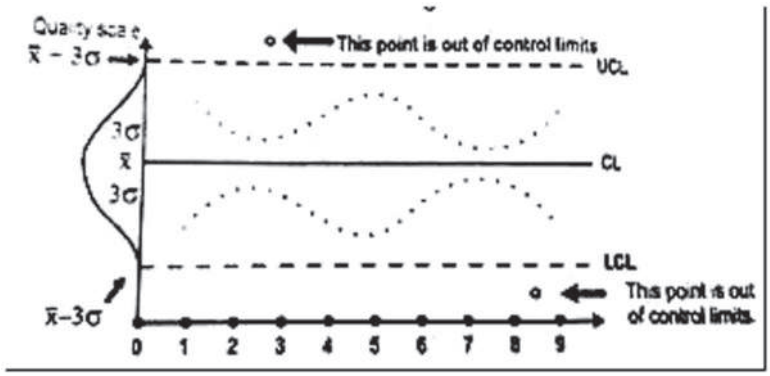
8.5 Statistical Quality Control Techniques

The quality of a product manufactured in a factory can be controlled in two stages. If the quality control mechanism is exercised during the processing period, the control is known as 'process control'. In this case, the quality is controlled when the product is at semi-finished stage. On the other hand, if the quality control mechanism is exercised after the completion of the production process, the control is known as 'product control'. In this case, the quality is controlled when the products are ready for sale. The statistical technique applied in process control is the control chart whereas in case of product control the acceptance plan is used.

8.6 Control Chart

A control chart is a graphical display of measurements of an industrial process through time. By carefully scrutinizing the chart, a quality control engineer can identify any potential problem associated with the production process. Dr. Walter A. Shewhart of Bell Laboratories, USA is the proponent of designing the control chart for industrial processes. At present, it is widely used in statistical quality control for maintaining the quality standard of the product produced in the production process. A control chart consists of three horizontal lines with a vertical line on the left. The vertical line represents the quality statistic of each sample and is known as quality scale. The base line of the chart shows the sample numbers and is used as subgroup scale. The horizontal line which passes through the middle of the chart parallel to the base line is called Central Line (CL). It indicates the desired standard of quality of the product in the process. The other two horizontal lines above and below the CL are known as Upper Control Limit (UCL) and Lower Control Limit (LCL) respectively. The UCL represents the upper limit of tolerance and the LCL indicates the lower limit of tolerance. The distances of the UCL and LCL from the CL are measured on the basis of probability. In most cases, the distance of the UCL from the CL and that of the LCL from the CL are both equal to 3 times the standard deviation of the sample characteristic for which the control chart in prepared. Thus these control limits taken together are also known as 'three-sigma control limits'. These limits provide the desired range of values for the statistic. When the statistic is outside the range of values, the process is considered to be out of control. The fact is immediately reported to the concerned authority who ascertains the causes of variation in the process and adopts appropriate corrective measures.

Control charts are based on normal distribution. The CL of the control chart is drawn at the mean $\bar{x}$ and the UCL and LCL pass through the chart three-sigma (3σ) above and 3σ below the CL respectively. In a normal distribution, the probability of a value lying between $\bar{x} \pm 3\sigma$ is 0.9973 and the probability of a point falling outside 3σ control limits is 0.0027. Thus, if the production process is under control, only about 27 out of 10000 events will fall outside the control limits. In other words, occurrence of events beyond the control limits is an extremely remote chance under normal circumstances. A specimen of the control chart is given below:



Sample (Sub-group) Number

The control chart technique can effectively be applied in statistical quality control for maintaining the quality standard of the product if only rational sub-groups are used. The rational sub-groups should be made by the efficient grouping of items in such a manner that variation in quality among items within the same group is small but variation between one group and another is as wide as possible.

8.7 Types of Control Chart

Control charts can be divided into two categories: (i) Control charts for variables and (ii) Control charts for attributes.

(i) Control charts for variables: Variables are those quantitative characters of a product which can be measured in terms of specific units of measurement, such as diameter of a pin, life of an electric bulb etc. In this case, for the purpose of quality control generally three types of control charts are used—

(a) Mean chart ($\bar{x}$ chart), (b) Range chart (R chart) and (c) Standard deviation chart (σ chart)

(ii) Control charts for attributes: Attributes are those qualitative characters of a product which cannot be measured in terms of specific units of measurement, but can only be identified by their presence or absence in the product. For example, we may say that glass is cracked or not cracked. In such cases, attributes may be assessed either by the proportion of units that are defective or by the number of defects per unit. Thus, for attributes two types of control charts are in common use—

(a) Fraction defective chart (ρ chart) and (b) Number of defects chart (c chart).

Let us now discuss various control charts in detail.

8.7.1 Mean Chart ($\bar{x}$ Chart)

Mean chart is prepared using rational sub-groups. In general, control limits are determined on the basis of smaller sub-groups of size 4 or 5 units and atleast 25 such sub-groups. The following steps are adopted at the time of preparing a mean chart:

i) Computation of the mean of each sub-group (sample): The mean of each sub-group (sample) is obtained by dividing the sum of the values included in the sample (sub-group) by the number of items (observations) in the sample (sub-group).

ii) Calculation of the mean of the sample means: The mean of the sample means is obtained by dividing the sum of the sample means by the number of samples (sub-groups) included in the chart. Suppose we have k samples each containing n observations. If $\bar{x}_i$ represents the mean of the i-th samples then the mean of the sample means is denoted by $\bar{\bar{x}}$

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + ... + \bar{x}_k}{k} = \frac{\sum \bar{x}_i}{k}$$

iii) Measurement of control limits: The upper control limit (UCL) and the lower control limit (LCL) can be ascertained by using the following formulae:

$$\text{UCL}_{\bar{x}} = \bar{\bar{X}} + 3\sigma_x$$

$$\text{LCL}_{\bar{x}} = \bar{\bar{X}} - 3\sigma_x$$

where σ_x represents the standard error (S.E.) of $\bar{x}$ based on all possible sample

means of size n and is equal to $\frac{\sigma}{\sqrt{n}}$. Here σ is estimated either by $\frac{\bar{\sigma}}{c_2}$ or by $\frac{\bar{R}}{d_2}$ where $\bar{\sigma}$ represents the average of sub group standard deviations and $\bar{R}$ represents the mean range (i.e. the average of sub-group ranges) and c_2 and d_2 are constants, the values of which being available in the tables showing conversion factors for control limits for different n.

For ascertaining the value of σ, the standard deviation of all process observations is used. However, this estimate is good only for large samples, n >10. When sample sizes are small, the mean range is used instead of $\bar{\sigma}$. Thus, the control limits will be—

Using standard deviation

$$\bar{\bar{X}} \pm 3\frac{\sigma}{\sqrt{n}}$$

$$= \bar{\bar{X}} \pm 3\left[\frac{\bar{\sigma}}{c_2}\right] \times \frac{1}{\sqrt{n}}$$

$$= \bar{\bar{X}} \pm \left[\frac{3}{c_2\sqrt{n}}\right]\bar{\sigma}$$

$$= \bar{\bar{X}} \pm A_1\bar{\sigma}$$

where $A_1 = \left[\frac{3}{c_2\sqrt{n}}\right]$

Using range

$$\bar{\bar{X}} \pm 3\frac{\sigma}{\sqrt{n}}$$

$$= \bar{\bar{X}} \pm 3\left[\frac{\bar{R}}{d_2}\right] \times \frac{1}{\sqrt{n}}$$

$$= \bar{\bar{X}} \pm \left[\frac{3}{d_2\sqrt{n}}\right]\bar{R}$$

$$= \bar{\bar{X}} \pm A_2\bar{R}$$

where $A_2 = \left[\frac{3}{d_2\sqrt{n}}\right]$

Therefore, using standard deviation $UCL_{\bar{X}} = \bar{\bar{X}} + A_1\bar{\sigma}$ and $LCL_{\bar{X}} = \bar{\bar{X}} - A_1\bar{\sigma}$ and using range $UCL_{\bar{X}} = \bar{\bar{X}} + A_2\bar{R}$ and $LCL_{\bar{X}} = \bar{\bar{X}} - A_2\bar{R}$.

iv) **Preparation of mean chart:** After determining control limits the mean chart is drawn on a graph paper.

Illustration 8.1

In order to construct a control chart, samples of size 3 are taken. The mean of the sample means is found to be 10.253 and the ranges of the samples of 3 observations each are 0.15, 0.53, 0.69, 0.45, 0.55, 0.71, 0.90, 0.68, 0.11 and 0.24. Find the upper and the lower control limits for mean chart. (Given that the constant of conversion for a sample of size n = 3 is A_2 =1.023).

Solution:

The following values are given:

n = 3, $\bar{\bar{x}}$ =10.253, A_2 = 1.023, R_1 = 0.15, R_2 = 0.53, R_3 = 0.69, R_4 = 0.45, R_5 = 0.55, R_6 = 0.71, R_7 = 0.90, R_8 = 0.68, R_9 = 0.11 and R_{10} = 0.24

$$\therefore R = \frac{(0.15+0.53+0.45+0.55+0.71+0.90+0.68+0.11+0.24)}{10} = 0.501$$

$$\therefore UCL_X = \bar{\bar{x}} + A_2\bar{R} = 10.253+(1.023)0.501=10.766$$

and $LCL_X = \bar{\bar{x}} - A_2\bar{R}$ = 10.253-(1.023)0.501 = 9.74

Illustration 8.2

A machine is set to deliver an item of a given weight. 10 samples of size 5 each were recorded. The relevant data are as follows:

Sample:	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{x}$):	15	17	15	18	17	14	18	15	17	16
Range (R):	7	7	4	9	8	7	12	4	11	5

(i) Ascertain the values of the CL, UCL and LCL for mean chart.

(ii) Comment on the state of control on the basis of the above values.

(Given: Conversion factors for n = 5 are A_2=0.58, D_3=0, D_4=2.115)

Solution:

From the given data, we have

$$\bar{\bar{x}} = \frac{\sum \bar{x}}{n} = \frac{162}{10} = 16.2 \text{ and}$$

$$\bar{R} = \frac{\sum R}{n} = \frac{74}{10} = 7.4$$

For the sample size, n=5, we know that A_2=0.58.

Thus, for the mean chart—

$$CL = \bar{\bar{x}} = 16.2$$

$$UCL = \bar{\bar{x}} + A_2\bar{R} = 16.2 + 0.58 \times (7.4) = 20.469$$

$$LCL = \bar{\bar{x}} - A_2\bar{R} = 16.2 - 0.58 \times (7.4) = 11.931$$

Since all the sample points fall within the control limits, the process is under statistical control.

Illustration 8.3

The following data relate to the life (in hours) of 7 samples of 6 electric bulbs each, drawn at intervals of one hour from a production process. Draw the chart and comment.

Sample No.	Life-time (in hours)					
1	520	582	565	655	729	580
2	401	483	425	490	560	570
3	562	601	585	472	521	553
4	545	525	472	527	534	645
5	395	885	569	542	550	530
6	532	652	527	482	583	445
7	529	610	570	593	662	434

(Given: Conversion factors for a sample of size n = 6 are A_2 = 0.483, D_3 = 0, D_4 = 2.004)

Solution:

For drawing $\bar{x}$ chart, we have to compute the mean and range for each of the sample.

Sample No.	Total	Mean	Range
1	3631	605.17	209
2	2929	488.17	169
3	3294	549.00	129
4	3248	541.33	173
5	3471	578.50	490
6	3221	536.83	207
7	3398	566.33	228
		3865.33	1605

$$\bar{\bar{x}} = \frac{3865.33}{7} = 552.19$$

and $$\bar{R} = \frac{1605}{7} = 229.29$$

n = 6 and A_2 = 0.483 (given)

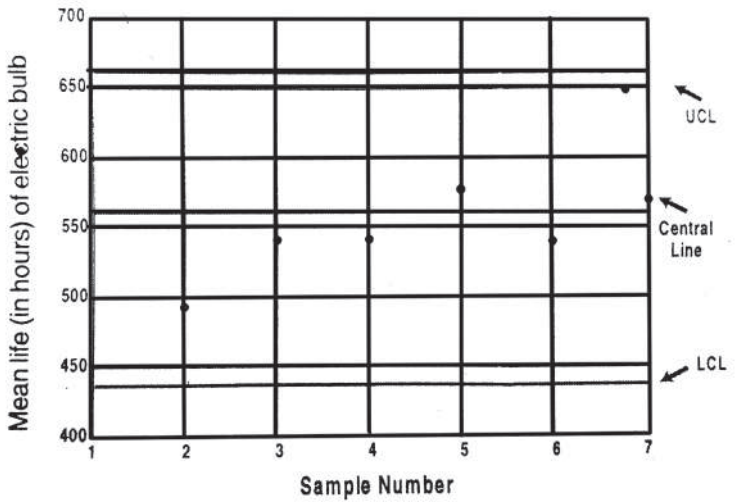
Thus, for the $\bar{x}$ chart,

$$UCL_{\bar{x}} = \bar{\bar{x}} + A_2\bar{R} = 552.19 + (0.483) \times 229.29 = 552.19 + 110.75 = 662.94$$

$$LCL_{\bar{x}} = \bar{\bar{x}} - A_2\bar{R} = 552.19 - (0.483) \times 229.29 = 552.19 - 110.75 = 441.44$$

Central line = $\bar{\bar{x}}$ =552.19

The $\bar{x}$ chart for life (in hours) of electric bulbs is drawn below:



The above $\bar{x}$ chart shows that all the sample means are well within the control limits. Thus, it implies that the process is in a state of control during the period under study so far as the process average is concerned.

8.7.2 Range Chart (R-Chart)

Range chart is prepared for the purpose of examining the variability of the quality produced by a given process specially when the sample sizes are small. The following steps are adopted in order to prepare a R-chart:

i) **Computation of range of each sample:** The range of each sample is computed. Let R_1, R_2, R_3, ……, R_k denote the values of corresponding ranges for the k samples. Thus for the i-th sample (i = 1,2,3, ……, k), we have the range $R_i=L_i - S_i$ where L_i represents the largest observation and S_i represents the smallest observation in the i-th sample.

ii) **Calculation of the mean of the sample ranges:** The mean of the sample ranges is obtained by dividing the sum of the sample ranges by the number of samples. Thus the mean of the sample ranges is—

$$\overline{R} = (R_1 + R_2 + R_3 + ... + R_k) \times \frac{1}{k} = \frac{\sum R_i}{k}$$

iii) Measurement of control limits: The upper control limit and the lower control limit are computed by using the following formulae:

$UCL_R = \overline{R} + 3\sigma_R$ and

$LCL_R = \overline{R} - 3\sigma_R$ where σ_R represents the standard error of the range.

The value of σ_R can be estimated by computing the standard deviation of the ranges of all possible samples of the size n drawn from a given population. But in practice it is convenient to determine the control limits by using the values of D_3 and D_4 which are available in the table showing conversion factors for control limits and central line for different n. Thus, when the tabulated values of D_3 and D_4 are used, the UCL and the LCL are determined by applying the following formulae:

$UCL_R = D_4\overline{R}$ and

$LCL_R = D_3\overline{R}$

The central line is $\overline{R}$. We know that range cannot be negative. Hence, if LCL_R comes out negative, then it is to be taken as zero.

iv) **Preparation of range chart:** After measuring the control limits the range chart is drawn on a graph paper.

Illustration 8.4

On the basis of the data given in Illustration 8.2 you are required to ascertain the values of the CL, UCL and LCL for the range chart and to comment on the state of control. Also interpret the results obtained from both Illustration 8.2 and Illustration 8.4

Solution:

For sample size, n = 5, we know that $D_3 = 0$ and $D_4 = 2.115$

CL = $\overline{R}$ = 7.4

$UCL = D_4\overline{R} = 2.115 \times 7.4 = 15.614$

$LCL = D_3\overline{R} = 0 \times 7.4 = 0$

Since all the sample points fall within the control limits, the process is under statistical control.

On the basis of the outcomes derived from the control limits for both $\bar{x}$-chart and R-chart, it can be concluded that the process is under control. Moreover, the process is free from any assignable causes of variation and is under the influence of only chance causes of variation.

Illustration 8.5

On the basis of the data given in Example 8.3 you are required to prepare R-chart and comment.

Solution: To examine whether the process dispersion is under control or not, we draw the range-chart.

$n = 6$, $D_3 = 0$ and $D_4 = 2.004$ (given)

$R_1 = 209$, $R_2 = 169$, $R_3 = 129$, $R_4 = 173$, $R_5 = 490$, $R_6 = 207$, $R_7 = 228$ and $\bar{R} = 229.29$ (computed).

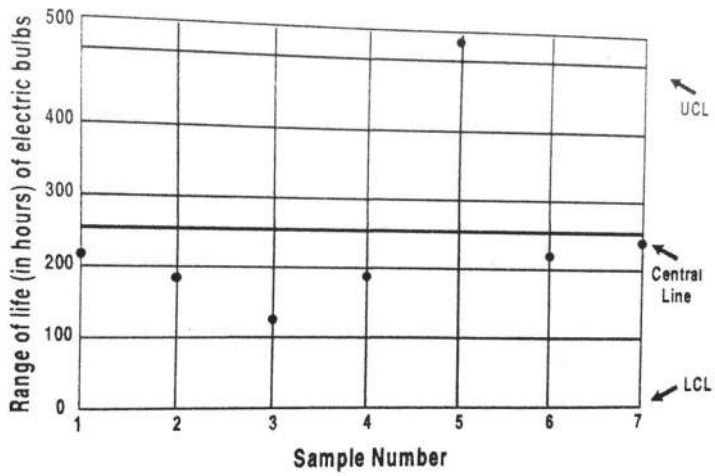
Thus, for the R-chart,

$$UCL_R = D_4 \bar{R} = 2.004 \times 229.29 = 459.50$$

$$LCL_R = D_3 \bar{R} = 0 \times 229.29 = 0$$

Central line = $\bar{R} = 229.29$

Thus R-chart for life (in hours) of electric bulbs is drawn below:



The above R-chart shows that all the sample ranges except the range for the firth sample are within the control limits. The range for the firth sample is slightly beyond the UCL. Thus, it implies that the process dispersion is almost in state of control although there is a sign of it going out of control in the fifth sample.

8.7.3 Control Chart For Standard Deviation (σ-Chart)

The R-chart is in common use in measuring the process variability because it is easier to compute ranges than to compute standard deviations. But standard deviation is considered to be an ideal measure of dispersion. Thus, σ-chart is theoretically more appropriate than R-chart for controlling the variability of the process. The control chart for the process standard deviation is similar to that for the range. While preparing σ-chart the constants B_3 and B_4 are used which are also found in the table showing conversion factors for control limits and central line. The control limits are measured by using the following formulae:

$UCL_\sigma = B_4 \bar{\sigma}$ and $LCL_\sigma = B_3 \bar{\sigma}$

The central line is $\bar{\sigma}$ where $\bar{\sigma}$ is the sum of sample standard deviations, divided by the number of samples. $\bar{\sigma}$-chart is interpreted exactly on the same lines as R-chart.

Illustration 8.6

In a cricket bat manufacturing unit the process quality was controlled by using control charts for mean and standard deviation. 20 samples of 15 items each were selected and then the sum of sample means and the sum of sample standard deviations were found to be 890.6 and 12.36 respectively. Determine the control limits and central line for σ-chart.

(Given: Conversion factors for a sample of size n=15 are B_3=0.428 and B_4=1.572)

Solution:

Given: n=15, B_3=0.428, B_4=1.572

Central line: $\bar{\sigma} = \frac{12.36}{20} = 0.618$

$UCL_\sigma = B_4\bar{\sigma} = 1.572 \times 0.618 = 0.9715$ (approx)

$LCL_\sigma = B_3\bar{\sigma} = 0.428 \times 0.618 = 0.2645$ (approx)

8.7.4 Fraction Defective Chart (ρ-Chart)

Fraction Defective Chart is applied in such cases where the quality characteristic of interest in an attribute and each manufactured item is classified into two categories only - defective and non-defective. In order to assess whether a process is in state of control or not, one will now try to ascertain if the population fraction defective is the same for all samples. The fraction defective in the sample (ρ) may be used as the basis of this test. The number of defective items in a random sample chosen from a population has binomial distribution. By the central limit theorem, as n increases, the distribution of the sample proportion (ρ) approaches a normal distribution. Thus, the fraction defective chart (ρ-chart) gives better results when the sample size is large. The central line of the ρ-chart is drawn at the average fraction defective ($\bar{\rho}$) from all the samples combined. The value of $\bar{\rho}$ is obtained by using the following formula:

$$\bar{\rho} = \frac{\text{Number of defective in all the samples combined}}{\text{Total number of items in all the samples combined}}$$

The upper and the lower limits of this chart are measured by applying the following formulae:

$$UCL_{\rho} = \bar{\rho} + 3\sqrt{\frac{\bar{\rho}\left(1-\bar{\rho}\right)}{n}}$$

$$LCL_{\rho} = \bar{\rho} - 3\sqrt{\frac{\bar{\rho}\left(1-\bar{\rho}\right)}{n}}$$

Example 8.7

The following figures give the number of defectives in 22 samples each containing 4000 items: 850, 860, 432, 682, 450, 644, 560, 612, 674, 610, 712, 804, 432, 528, 252, 818, 386, 652, 560, 778, 902, 840.

Calculate the values for central line and control limits for ρ-chart and draw the chart. Also comment on the state of control of the process.

Solution:

The total number of defectives out of (22 × 4000) or 88000 items inspected in the 22 samples is:

Σd = (850+860+432+682+450+644+560+612+674+610+712+804+432+528+ 252+818+386+652+560+778+902+840) = 14,038

The average fraction defective is

$$\bar{\rho} = \frac{14038}{22 \times 4000} = 0.1595 \text{ (approx.)}$$

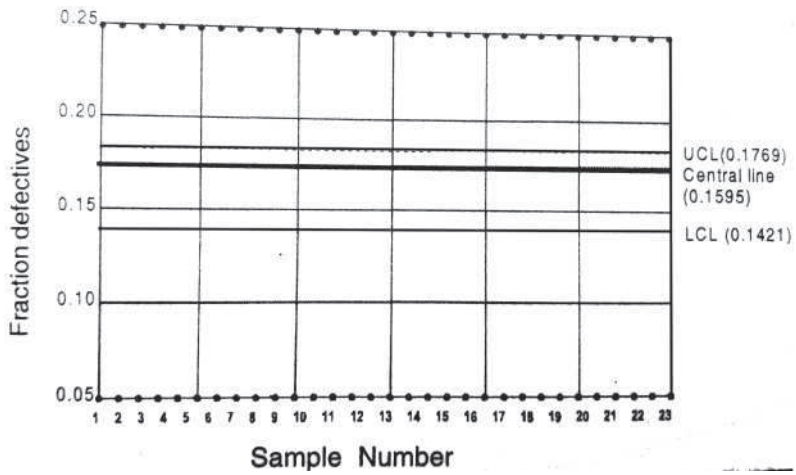
For ρ-chart:

$$UCL_{\rho} = \bar{\rho} + 3\sqrt{\frac{\bar{\rho}(1-\bar{\rho})}{n}} = 0.1595 + 3\sqrt{\frac{0.1595(1-0.1595)}{4000}} = 0.1769$$

$$LCL_{\rho} = \bar{\rho} - 3\sqrt{\frac{\bar{\rho}(1-\bar{\rho})}{n}} = 0.1595 - 3\sqrt{\frac{0.1595(1-0.1595)}{4000}} = 0.1421$$

And central line = $\bar{\rho}$ = 0.1595

The ρ-chart is drawn below—



The above ρ-chart shows that quite a large number of dots are outside the control limits. It implies that the production process is completely out of control.

8.7.5 Control Chart for Number of Defects (c-Chart)

Control chart for number of defects is prepared for the purpose of exercising the number of defects or imperfections per item. For example, when glass bottles are manufactured, it is of interest to keep a record of the number of air bubbles per bottle and to adopt corrective measure if this number is out of control. The number of

defects, denoted by c, may in most cases be supposed to have a Poisson distribution. For the Poisson distribution, we know that the mean and the variance are both equal to the same parameter. The Poisson distribution can be approximated by the normal distribution for large sample sizes. Thus, the central line of the c-chart is $\bar{c}$ and the control limits of the chart are:

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}}$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}}$$

where $\bar{c}$ is the average number of defects or imperfections per item. We know that $\bar{c}$ cannot be negative. Thus, in the c-chart if the formula for LCL_C yields negative value, then it is to be considered as zero.

Illustration 8.8

The following are the number of defects observed in 20 hundred-yard pieces of woolen goods:

5, 5, 8, 5, 0, 2, 5, 8, 10, 11, 6, 14, 8, 8, 6, 5, 6, 7, 2, 2

Find $\bar{c}$ and the control limits for the number of defects.

Solution:

Total number of defects $\Sigma c = 123$

$\therefore$ Average number of defects per piece $(\bar{c}) = \frac{123}{20} = 6.15$

For c-chart: Central line = $\bar{c} = 6.15$

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}} = 6.15 + 3\sqrt{6.15} = 6.15 + 7.44 = 13.59$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}} = 6.15 - 3\sqrt{6.15} = 6.15 - 7.44, \text{ which is negative}$$

$\therefore$ $LCL_C = 0$

8.8 Acceptance Sampling

While the control charts are used in process control, the acceptance sampling technique is applied in product control. When finished products are presented by the manufacturer to the consumer for acceptance, the latter intends to examine whether the product conforms to specifications or not. Such examination is done either on the basis of 'complete enumeration' or on the basis of 'sampling'. In most of the cases, sampling inspection technique is followed because - (i) it takes less time, labour and money and (ii) it also creates effective pressure for quality improvement. The use of sampling inspection by a consumer for taking decision on acceptance or rejection of the product is known as 'acceptance sampling'. It may be of two types - (i) Lot-by-lot sampling inspection and (ii) Continuous sampling inspection. In the lot-by-lot sampling inspection, the finished articles are formed into lots, a few items are chosen randomly and the lot is either accepted or rejected on the basis of a certain set of sampling rules, usually called 'Sampling Inspection Plan'. In the sampling inspection plan if the quality is measured in terms of a variable, then it is known as 'Variable Sampling Inspection Plan' and if the quality is measured in terms of an attribute, then it is known as 'Attribute Sampling Inspection Plan'. In the continuous sampling inspection, the outcome of current inspection is considered as the basis of ascertaining the need for total inspection or sampling inspection for the next batch of manufactured articles to be inspected.

8.9 Attributes Sampling Inspection Plans

There are three types of attribute sampling inspection plans - 'Single Sampling Inspection Plan', 'Double Sampling Inspection Plan' and 'Multiple Sampling Inspection Plan'.

Let us now discuss these three plans one by one.

8.9.1 Single Sampling Inspection Plan

If the decision whether to accept or reject a lot is made on the basis of only one sample drawn randomly from the lot, then the sampling inspection plan is known as the single sampling inspection plan. Let N and n represent lot size and sample size respectively. Let c denote acceptance number, i.e. the maximum number of defective articles allowable in the sample and let d represent the actual number of defectives in the sample. Then the single sampling inspection plan involves the following steps:

i) A sample of n items is selected randomly from a lot of N items.

ii) The defective items (d) are found out by inspecting the sample thoroughly.

iii) If $d \leq c$, then the lot is accepted, but if $d>c$, then the lot is rejected.

Let us explain it with the help of the following example.

Illustration 8.9

Consider the following single sampling plan:

$$N = 600,\ n = 45,\ c = 6$$

Interpret these numbers.

Solution:

The given numbers can be interpreted as follows:

A random sample of 45 items from a lot containing 600 items is taken. If the sample contains at most 6 defective items, then the lot will be accepted. If the number of defective items exceeds 6, the lot will be rejected.

8.9.2 Double Sampling Inspection Plan

If the decision whether to accept or reject a lot is made on the basis of two samples, then the sampling inspection plan is known as the double sampling inspection plan.

Let N, n_1, n_2 represent lot size, size of the first sample and size of the second sample respectively. Let c_1 and c_2 denote acceptance number of the first sample and acceptance number for the two samples combined respectively and let d_1 and d_2 represent the numbers of defectives in the first and second samples respectively.

Then the double sampling inspection plan involves the following steps:

i) A sample of n_1 items is first selected at random from the lot of size N.

ii) The value of d_1 is ascertained by a careful inspection of the sample.

iii) If $d_1 \leq c_1$, then the lot is accepted and if $d_1>c_2$, then the lot is rejected.

iv) If $d_1>c_1$, but $d_1 \leq c_2$ (i.e. $d_1 \ngtr c_2$), a second sample of n_2 items is chosen from the same lot.

v) If $(d_1+d_2) \leq c_2$, the lot is accepted, otherwise it is rejected.

Let us explain the plan with the help of the following example.

Illustration 8.10

Consider the following double sampling inspection plan:

$$N = 6000,\ n_1 = 300,\ c_1 = 6,\ n_2 = 240,\ c_2 = 18$$

Interpret these numbers.

Solution:

The given numbers can be interpreted as follows:

i) A random sample of 300 items from a lot containing 6000 items is taken.

ii) If the sample is found to contain at most 6 defective items, the lot is accepted; if the sample contains more than 18 defective items, the lot is rejected.

iii) If the number of defectives in the first sample (d_1) exceeds 6 but does not exceed 18, a second sample of 240 items is chosen.

iv) If the combined sample ($n_1 + n_2$) of 540 items contains at most 18 defectives, the lot is accepted and if the combined sample contains more than 18 defective items, the lot is rejected.

8.9.3 Multiple Sampling Inspection Plan

If the decision whether to accept or reject a lot is made on the basis of more than two samples, then the sampling inspection plan is known as the multiple sampling inspection plan. It is quite complicated and rarely used in practice.

8.10 Summary

Statistical Quality Control (SQC) is the statistical techniques used to maintain uniform quality of products in a continuous flow of manufacturing process. The variation in the quality of products is mainly due to two distinct types of causes - chance causes and assignable causes. The quality can be controlled either in the processing period or after 'acceptance plan' is used. Control charts are mainly divided into categories - control charts for variables and control charts for attributes. The most common control charts for variables are mean chart ($\bar{x}$-chart), range chart (R-chart) and standard deviation chart (σ-chart) whereas for attributes the major ones are fraction chart (ρ-chart) and the number of defects chart (c-chart). In order to examine whether the product conforms to specifications or not, some form of inspection either on the basis of 'complete enumeration' or on the basis of 'sampling' is needed. In most

cases, sampling inspection technique is followed. Sampling inspection may be of two types - lot-by-lot sampling inspection and continuous sampling inspection. Further, it may be either variable sampling inspection or attribute sampling inspection. There are three types of attribute sampling inspection plans - single sampling inspection plan, double sampling inspection plan and multiple sampling inspection plan.

8.11 Self-Assessment Questions

Long Answer Type Questions

1. What is Statistical Quality Control? Discuss the need for Statistical Quality Control.
2. "Variations in the quality of products in a manufacturing process are attributed to two distinct types of causes". Explain.
3. What do you mean by the term 'rational sub-group'? Distinguish between product control and process control.
4. What are control charts? State the different types of control charts for variables and attributes.
5. What is acceptance sampling? Why is it used? State the conditions for its use.
6. Narrate the different types of sampling inspection plan popularly used in practice.
7. The following table discloses the length of tin-plate produced in a machine, as obtained from 8 samples of size 6 each drawn at regular intervals of 20 minutes:

Sample No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	23	24	20	25	26	24	15	13
2	17	21	19	21	22	17	9	17
3	22	23	16	18	19	21	16	19
4	19	20	19	17	16	18	17	22
5	12	14	12	13	21	20	23	16
6	11	16	15	9	20	10	20	24

Construct (i) a mean chart and (ii) a range chart.

Also comment on the state of control.

(Conversion factors for n=6 are A_2=0.483, D_3=0, D_4=2.004)

8. Using the data given in the problem no. 7, calculate the values for control line and the control limits for σ-chart and draw your conclusions:

   (Given: Conversion factors for n=6 are B3=0.03 and B4=1.97).

9. The following table shows the number of defects observed in 10 woolen carpets passing as satisfactory:

Serial no of carpets:	1	2	3	4	5	6	7	8	9	10
No of defects:	3	4	5	5	1	6	2	2	7	4

   Construct c-chart.

10. The following table shows the numbers of missing nuts observed at the time of final inspection of 15 machines:

Machine No:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of missing nuts:	3	10	2	5	6	4	9	1	7	2	8	6	6	7	8

    Find the control limits for the number of defects chart and comment on the state of control.

Short Answer Type Questions

1. Write short notes on:

   a) Mean chart

   b) Range chart

   c) Control chart for standard deviation

   d) Fraction defective chart

   e) Control chart for number of defects.

2. In order to construct a control chart, samples of size 28 are taken. The mean sample range is found to be 0.32 and the process average of the machine is set at 17.21. Find the lower and the upper control limits for mean. (Given that the constant of conversion for a sample of size n=25 is A_2=0.153).

3. Determine control limits on a mean chart for samples of size 6 if the process has to meet a lower and an upper specification limits of 159 cms. and 295 cms. respectively.

4. Based on 30 samples each of size 300 taken at intervals of 1 hour from a manufacturing process, the average fraction defective was found to be 0.073. Determine the values of central line and the control limits for a ρ-chart.
5. The number of defectives in 15 samples (each containing 1000 items) are given below:

   125, 230, 170, 168, 190, 120, 250, 225, 200, 210, 170, 160, 180, 190, 205

   Calculate the values for central line and control limits for ρ-chart.

Objective Answer Type Questions

1. What is SQC?
2. What do you mean by chance causes in the context of SQC?
3. Define 'assignable causes' in the context of SQC?
4. What is control chart?
5. What is mean chart?
6. What is range chart?
7. What is control chart for standard deviation?
8. What is ρ-chart?
9. What is c-chart?
10. What do you mean by acceptance sampling?

Suggested Readings

1. A. M. Goon, M.K. Gupta, B. Dasgupta, Fundamentals of statistics, Vol-I. The world Press Private Limited.
2. Arora et.al. (2007): Comprehensive Statistical Methods, S. Chand & Co. Ltd., New Delhi.
3. C. R. Kothari, Research Methodology, Wishwa Prakashan.
4. Chandra (2001): Statistical Quality Control, CRC Press, USA
5. Das (1984): Statistical Methods in Commerce, Accounting & Economics (Part II), M. Das & Co., Kolkata
6. D. R. Anderson, D. J. Sweeney, T. A. Williams, Statistics for Business and Economics, Cengage Learning.
7. Gupta (2007): Fundamentals of Statistics, Himalaya Publishing House, Mumbai.
8. Gupta & Gupta (2010): An Introduction to Statistical Methods, Vikas Publishing House Pvt. Ltd., New Delhi.
9. Gupta & Kapoor (1997): Fundamentals of Mathematical Statistics, Sultan Chand & Sons, New Delhi.
10. Kothari (2002): Research Methodology—Methods & Techniques, Wishwa Prakashan, New Delhi.
11. N. G. Das, Statistical Methods, Vol-II
12. R. I. Levin, D.S. Rubin, statistics for Mnanagement.
13. S. B. Choudhury, Elementary statistics, Vol-I, The world press private Limited.

Appendix Table 1 : Area Under standard Normal Curve

(The given proportions indicate area above the given value of Z)

Normal Deviate Z	.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4062	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2258	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1822	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0156	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0033	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

Appendix Table 2: Percentile Values of the Student's t- distribution

d f \ 1– α	0.75	0.9	0.95	0.975	0.99	0.995	0.9995
1	1	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.886	2.92	4.303	6.965	9.925	31.598
3	0.765	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.61
5	0.727	1.474	2.015	2.571	3.365	4.032	6.859
6	0.718	1.44	1.913	2.447	3.143	3.707	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	1.387	1.86	2.306	2.896	3.355	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	1.356	1.782	2.179	2.671	3.055	4.318
13	0.694	1.35	1.771	2.160	2.65	3.012	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	1.337	1.746	2.12	2.583	2.921	4.015
17	0.689	1.333	1.740	2.11	2.567	2.898	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.992
19	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	1.323	1.721	2.08	2.518	2.831	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.699	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.291

Appendix Table 3 : Percentile Values of the the Chi-square distribution

d f \ α	0.99	0.98	0.95	0.9	0.8	0.7	0.5	0.05	0.01	0.001
1	...	...	0.00393	0.0158	0.642	0.148	0.455	3.841	6.635	10.827
2	0.0201	0.0401	0.103	0.211	0.446	0.713	2.386	5.991	9.21	13.815
3	0.115	0.185	0.352	0.584	1.005	1.414	2.366	7.815	11.345	16.266
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	9.488	13.277	18.467
5	0.554	0.752	1.145	1.710	2.343	3.000	4.351	11.07	15.086	20.515
6	0.862	1.134	1.635	2.204	3.07	3.828	5.348	12.592	16.812	22.457
7	1.139	1.564	2.167	2.833	3.822	4.671	6.346	14.067	18.475	24.322
8	1.646	2.032	2.733	3.49	4.594	5.527	7.344	15.507	20.09	26.125
9	2.088	2.532	3.325	4.168	5.38	6.393	8.343	16.919	21.666	27.877
10	2.558	3.059	3.94	4.865	6.179	7.267	9.342	18.307	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	19.675	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.024	11.34	21.026	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.34	22.362	27.688	34.528
14	4.66	5.368	6.571	7.79	9.467	10.821	13.339	23.685	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	24.996	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	26.296	32.001	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	27.587	33.409	40.791
18	7.015	7.906	9.390	10.865	12.857	14.44	17.338	28.869	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	30.144	36.191	43.821
20	8.26	9.237	10.851	12.443	14.578	16.266	19.337	31.41	36.566	45.315
21	8.897	9.915	11.591	13.24	15.445	17.182	20.337	32.641	38.932	46.798
22	9.542	10.6	12.238	14.041	16.314	18.101	21.337	33.924	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.201	22.337	35.172	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	36415	42.98	51.179
25	11.524	12.697	14.611	16.473	18.94	20.867	24.337	37.652	44.314	52.620
26	12.198	13.409	15.379	17.292	19.82	21.792	25.336	38.885	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	40.113	46.963	55.467
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	41.337	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	42.557	49.588	58.303
30	14.953	16.306	18.493	20.589	23.364	25.508	29.336	43.773	50.892	59.703
40	22.164	23.838	26.509	29.051	32.345	34.872	39.335	55.759	63.692	73.402
50	29.707	31.644	34.764	37.689	41.449	44.313	34.335	67.595	76.154	86.661
60	37.485	39.699	43.188	46.459	50.641	53.809	59.335	79.082	88.379	99.607

Appendix Table 4: 5% percent values of the F-distribution

$df(n_2)$	$df(n_1)$ 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234	236.8	283.9	240.5	241.9	243.9	245.9	248	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.4	19.41	19.43	19.45	19.45	1946	19.47	19.48	19.49	19.5
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.7	8.66	8.64	8.62	8.59	8.57	8.55	8.5
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.8	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.5	4.46	4.43	4.4	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4	3.94	3.87	3.84	3.81	3.77	3.74	3.7	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.3	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.9	2.86	2.83	2.79	2.75	2.71
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.7	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.4
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.3
13	4.64	3.81	3.46	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.6	2.53	2.46	2.42	2.38	2.34	2.3	2.25	2.21
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.48	2.4	2.33	2.29	2.25	2.2	2.16	2.11	2.07
16	4.49	3.63	3.34	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.1	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.28	2.2	2.12	2.08	2.04	1.99	1.95	1.9	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.1	2.05	2.01	1.96	1.92	1.87	1.81
22	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34	2.3	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32	2.27	2.2	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.9	1.85	1.8	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.2	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.2	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.7	2.55	2.43	2.35	2.28	2.22	2.18	2.1	2.03	1.94	1.9	1.85	1.81	1.75	1.7	1.64
30	4.17	3.32	2.92	2.69	2.45	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.37	2.34	2.25	2.18	2.12	2.08	2	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.01	3.15	2.76	2.53	2.29	2.25	2.17	2.1	2.04	1.99	1.92	1.84	1.75	1.7	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.43	2.21	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.5	1.43	1.35	1.25
∞	3.84	3	2.6	2.37	2.19	2.1	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	8

Appendix Table 5: 1% percent values of the F-distribution

	d f (n_1)																		
d f (n_2)	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	....
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,022	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.17	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.50	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
.....	6.63	4.61	9.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Appendix Table 6 : Random Numbers

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	2315	7548	5901	8372	5993	7624	9708	8695	2303	6744
2	0554	5550	4310	5374	3508	9061	1837	4410	9622	1343
3	1487	1603	5082	4043	6223	5005	1003	2211	5438	0834
4	3897	6749	5194	0517	5833	7880	5901	9432	4287	1695
5	9731	2617	1899	7553	0870	9425	1258	4154	8821	0513
6	1174	2693	8144	3393	0872	3279	7331	1822	6470	6850
7	4336	1288	5911	0164	5623	9300	9004	9943	6407	4036
8	9380	6204	7838	2680	4491	5565	1189	3258	4755	2571
9	4954	0131	8108	4298	4187	6953	8296	6177	7380	9527
10	3676	8726	3337	9482	9569	4195	9686	7045	2748	3880
11	0709	2523	9224	6271	2607	0655	8453	4467	3384	5320
12	4331	0010	8144	8638	0307	5255	5161	4889	7429	4647
13	6157	0053	6006	1736	3775	6314	8951	2335	0174	6993
14	3135	2837	9910	7791	8941	3157	9764	4862	5848	6919
15	5704	8865	2627	7959	3682	9052	9565	4635	0653	2254
16	0924	3442	0068	7210	7137	3072	9757	5609	2982	7650
17	9795	5350	1840	8948	8329	5223	0825	2122	5326	1587
18	9373	2595	7043	7819	8885	5667	1668	3695	9964	4569
19	7262	1112	2500	9226	8264	3566	6594	3471	6875	1867
20	6102	0744	1845	3712	0794	9511	7378	6699	5361	9378
21	9783	9854	7433	0559	1718	4547	3541	4422	0342	3000
22	8916	0971	9222	2329	0637	3505	5454	8988	4381	6361
23	2596	6882	2062	8717	9265	0292	3528	6248	9195	4883
24	8144	2317	1905	0495	4806	7569	0075	6765	0171	6545
25	1132	2549	3142	3623	4386	0862	4976	6762	2452	3245

Notes

Notes

Notes