STATISTICS

HIGHER SECONDARY - SECOND YEAR

Untouchability is a sin

Untouchability is a crime

Untouchability is inhuman

TAMILNADU
TEXTBOOK CORPORATION
College Road, Chennai- 600 006

First Edition – 2005
Reprint – 2006

Chairperson
Dr. J. Jothikumar
Reader in Statistics
Presidency College
Chennai - 600 005.

Reviewers and Authors

Thiru K.Nagabushanam
S.G.Lecturer in Statistics
Presidency College
Chennai - 600 005.

Dr. R.Ravanan
Reader in Statistics
Presidency College
Chennai - 600 005.

Authors

Thiru G.Gnana Sundaram
P.G.Teacher
S.S.V. Hr. Sec. School
Parktown, Chennai - 600 003.

Tmt.N.Suseela
P.G.Teacher
Anna Adarsh Matric HSS
Annanagar, Chennai -600 040

Tmt.S.Ezhilarasi
P.G.Teacher
P.K.G.G. Hr. Sec. School
Ambattur, Chennai -600 053.

Thiru A. S. Sekar
P.G.Teacher
O.R.G.N. Govt Boys HSS
Redhills, Chennai - 600 052

Price: Rs.

This book has been prepared by the Directorate of School Education on behalf of the Government of Tamilnadu.

This book has been printed on 60 G.S.M paper

Printed by offset at:

PREFACE

We take great pleasure in presenting the book on Statistics to the students of the Second year Higher Secondary classes.

This book has been written in conformity with the revised syllabus. The book is designed to be self-contained and comprises of ten chapters and includes two new chapters Association of attributes and Decision Theory. Besides the additional (new) topics covered in this book, all the chapters have been completely rewritten and simplified in many ways.

The book covers the theoretical, practical and applied aspects of statistics as far as possible in a clear and exhaustive manner. Every chapter in this book explains the principles through appropriate examples in a graded manner. A set of exercise concludes each chapter to provide an opportunity for the students to reinforce what they learn, to test their progress and increase their confidence.

The book is very helpful to the students who take their higher studies and the professional courses like Charted Accountants and ICWA

At the end of this textbook, necessary statistical tables are included for the convenience of the students.

We welcome suggestions from students, teachers and academicians so that this book may further be improved upon.

We thank everyone who has a helping hand in the lent preparation of this book.

Dr. J. Jothikumar
Chairperson

Writing team

CONTENTS

1. PROBABILITY

1.0 Introduction:

The theory of probability has its origin in the games of chance related to gambling such as tossing of a coin, throwing of a die, drawing cards from a pack of cards etc. Jerame Cardon, an Italian mathematician wrote ' A book on games of chance' which was published on 1663. Starting with games of chance, probability has become one of the basic tools of statistics. The knowledge of probability theory makes it possible to interpret statistical results, since many statistical procedures involve conclusions based on samples.

Probability theory is being applied in the solution of social, economic, business problems. Today the concept of probability has assumed greater importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision-making research. Probability theory, in fact, is the foundation of statistical inferences.

1.1 Definitions and basic concepts:

The following definitions and terms are used in studying the theory of probability.

Random experiment:

Random experiment is one whose results depend on chance, that is the result cannot be predicted. Tossing of coins, throwing of dice are some examples of random experiments.

Trial:

Performing a random experiment is called a trial.

Outcomes:

The results of a random experiment are called its outcomes. When two coins are tossed the possible outcomes are HH, HT, TH, TT.

Event:

An outcome or a combination of outcomes of a random experiment is called an event. For example tossing of a coin is a random experiment and getting a head or tail is an event.

Sample space:

Each conceivable outcome of an experiment is called a sample point. The totality of all sample points is called a sample space and is denoted by **S**. For example, when a coin is tossed, the sample space is $S = \{ H, T \}$. H and T are the sample points of the sample space S.

Equally likely events:

Two or more events are said to be equally likely if each one of them has an equal chance of occurring. For example in tossing of a coin, the event of getting a head and the event of getting a tail are equally likely events.

Mutually exclusive events:

Two or more events are said to be mutually exclusive, when the occurrence of any one event excludes the occurrence of the other event. Mutually exclusive events cannot occur simultaneously.

For example when a coin is tossed, either the head or the tail will come up. Therefore the occurrence of the head completely excludes the occurrence of the tail. Thus getting head or tail in tossing of a coin is a mutually exclusive event.

Exhaustive events:

Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment. For example, while throwing a die, the possible outcomes are {1, 2, 3, 4, 5, 6} and hence the number of cases is 6.

Complementary events:

The event ‘ A occurs’ and the event ‘ A does not occur’ are called complementary events to each other. The event ‘ A does not occur’ is denoted by A' or $\overline{A}$ or A^c. The event and its complements are mutually exclusive. For example in throwing a die, the event of getting odd numbers is { 1, 3, 5 } and getting even numbers is

{2, 4, 6}.These two events are mutually exclusive and complement to each other.

Independent events:

Events are said to be independent if the occurrence of one does not affect the others. In the experiment of tossing a fair coin, the occurrence of the event ‘ head’ in the first toss is independent of the occurrence of the event ‘ head’ in the second toss, third toss and subsequent tosses.

1.2 Definitions of Probability:

There are two types of probability. They are Mathematical probability and Statistical probability.

1.2.1 Mathematical Probability (or a priori probability):

If the probability of an event can be calculated even before the actual happening of the event, that is, even before conducting the experiment, it is called ***Mathematical probability.***

If the random experiments results in *‘n’* exhaustive, mutually exclusive and equally likely cases, out of which *‘m’* are favourable to the occurrence of an event A, then the ratio *m/n* is called the probability of occurrence of event A, denoted by P(A), is given by

$$P(A) = \frac{m}{n} = \frac{\text{Number of cases favourable to the event A}}{\text{Total number of exhaustive cases}}$$

Mathematical probability is often called ***classical probability*** or a ***priori probability*** because if we keep using the examples of tossing of fair coin, dice etc., we can state the answer in advance (*prior*), without tossing of coins or without rolling the dice etc.,

The above definition of probability is widely used, but it cannot be applied under the following situations:

(1) If it is not possible to enumerate all the possible outcomes for an experiment.

(2) If the sample points(outcomes) are not mutually independent.

(3) If the total number of outcomes is infinite.

(4) If each and every outcome is not equally likely.

Some of the drawbacks of classical probability are removed in another definition given below:

1.2.2 Statistical Probability (or a posteriori probability):

If the probability of an event can be determined only after the actual happening of the event, it is called ***Statistical probability.***

If an event occurs ***m*** times out of ***n***, its relative frequency is ***m/n.***

In the limiting case, when ***n*** becomes sufficiently large it corresponds to a number which is called the probability of that event.

In symbol, $P(A) = \lim_{n\to\infty} (m/n)$

The above definition of probability involves a concept which has a long term consequence. This approach was initiated by the mathematician Von Mises .

If a coin is tossed 10 times we may get 6 heads and 4 tails or 4 heads and 6 tails or any other result. In these cases the probability of getting a head is **not 0.5** as we consider in Mathematical probability.

However, if the experiment is carried out a large number of times we should expect approximately equal number of heads and tails and we can see that the probability of getting head approaches 0.5. The Statistical probability calculated by conducting an actual experiment is also called a ***posteriori probability*** or ***empirical probability.***

1.2.3 Axiomatic approach to probability:

The modern approach to probability is purely axiomatic and it is based on the set theory. The axiomatic approach to probability was introduced by the Russian mathematician A.N. Kolmogorov in the year 1933.

Axioms of probability:

Let S be a sample space and A be an event in S and P(A) is the probability satisfying the following axioms:

(1) The probability of any event ranges from zero to one.
i.e $0 \le P(A) \le 1$

(2) The probability of the entire space is 1.
i.e $P(S) = 1$

(3) If A_1, A_2,…is a sequence of mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

Interpretation of statistical statements in terms of set theory:

$S \Rightarrow$ Sample space

$\overline{A} \Rightarrow$ A does not occur

$A \cup \overline{A} = S$

$A \cap B = \phi \Rightarrow$ A and B are mutually exclusive.

$A \cup B \Rightarrow$ Event A occurs or B occurs or both A and B occur. (at least one of the events A or B occurs)

$A \cap B \Rightarrow$ Both the events A and B occur.

$\overline{A} \cap \overline{B} \Rightarrow$ Neither A nor B occurs

$A \cap \overline{B} \Rightarrow$ Event A occurs and B does not occur

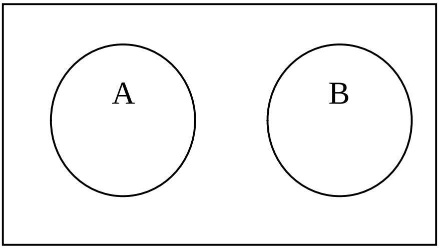
$\overline{A} \cap B \Rightarrow$ Event A does not occur and B occur.

1.3 Addition theorem on probabilities:

We shall discuss the addition theorem on probabilities for mutually exclusive events and not mutually exclusive events.

1.3.1 Addition theorem on probabilities for mutually exclusive events:

If two events A and B are mutually exclusive, the probability of the occurrence of either A or B is the sum of individual probabilities of A and B. ie $P(AUB) = P(A) + P(B)$
This is clearly stated in axioms of probability.



1.3.2 Addition theorem on probabilities for not-mutually exclusive events:

If two events A and B are not-mutually exclusive, the probability of the event that either A or B or both occur is given as

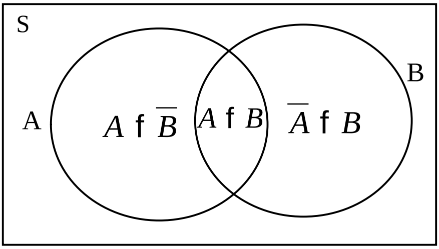
$$P(AUB) = P(A) + P(B) – P(A \cap B)$$

Proof:

Let us take a random experiment with a sample space S of N sample points.

Then by the definition of probability ,

$$P(AUB) = \frac{n(AUB)}{n(S)} = \frac{n(AUB)}{N}$$



From the diagram, using the axiom for the mutually exclusive events, we write

$$P(AUB) = \frac{n(A)+n(\overline{A} \cap B)}{N}$$

Adding and subtracting n($A \cap B$) in the numerator,

$$= \frac{n(A)+n(\overline{A} \cap B)+n(A \cap B)-n(A \cap B)}{N}$$

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

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

$$P(AUB) = P(A) + P(B) – P(A \cap B)$$

Note:

In the case of three events A,B,C, P(AUBUC) = P(A) + P(B) + P(C) – P(A f B) – P(A f B) – P(B f C) + P (A f B f C)

Compound events:

The joint occurrence of two or more events is called compound events. Thus compound events imply the simultaneous occurrence of two or more simple events.

For example, in tossing of two fair coins simultaneously, the event of getting ‘ atleast one head’ is a compound event as it consists of joint occurrence of two simple events.

Namely,

Event A = one head appears ie A = { HT, TH} and

Event B = two heads appears ie B = {HH}

Similarly, if a bag contains 6 white and 6 red balls and we make a draw of 2 balls at random, then the events that ‘ both are white’ or one is white and one is red’ are compound events.

The compound events may be further classified as

(1) Independent event

(2) Dependent event

Independent events:

If two or more events occur in such a way that the occurrence of one does not affect the occurrence of another, they are said to be independent events.

For example, if a coin is tossed twice, the results of the second throw would in no way be affected by the results of the first throw.

Similarly, if a bag contains 5 white and 7 red balls and then two balls are drawn one by one in such a way that the first ball is replaced before the second one is drawn. In this situation, the two events, ‘ the first ball is white’ and ‘ second ball is red’ , will be independent, since the composition of the balls in the bag remains unchanged before a second draw is made.

Dependent events:

If the occurrence of one event influences the occurrence of the other, then the second event is said to be dependent on the first.

In the above example, if we do not replace the first ball drawn, this will change the composition of balls in the bag while making the second draw and therefore the event of ‘ drawing a red ball’ in the second will depend on event (first ball is red or white) occurring in first draw.

Similarly, if a person draw a card from a full pack and does not replace it, the result of the draw made afterwards will be dependent on the first draw.

1.4 Conditional probability:

Let A be any event with p(A) >0. The probability that an event B occurs subject to the condition that A has already occurred is known as the conditional probability of occurrence of the event B on the assumption that the event A has already occurred and is denoted by the symbol P(B/A) or P(B|A) and is read as the probability of B given A.

The same definition can be given as follows also:

Two events A and B are said to be dependent when A can occur only when B is known to have occurred (or vice versa). The probability attached to such an event is called the **conditional probability** and is denoted by P(B/A) or, in other words, probability of B given that A has occurred.

If two events A and B are dependent, then the conditional probability of B given A is

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

Similarly the conditional probability of A given B is given as

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

Note:

If the events A and B are independent, that is the probability of occurrence of any one of them P(A/B) = P(A) and P(B/A) = P(B)

1.5 Multiplication theorem on probabilities:

We shall discuss multiplication theorem on probabilities for both independent and dependent events.

1.5.1 Multiplication theorem on probabilities for independent events:

If two events A and B are independent, the probability that both of them occur is equal to the product of their individual probabilities. i.e P(A f B) = P(A) .P(B)

Proof:

Out of n_1 possible cases let m_1 cases be favourable for the occurrence of the event A.

$$\therefore P(A) = \frac{m_1}{n_1}$$

Out of n_2 possible cases, let m_2 cases be favourable for the occurrence of the event B

$$\therefore P(B) = \frac{m_2}{n_2}$$

Each of n_1 possible cases can be associated with each of the n_2 possible cases.

Therefore the total number of possible cases for the occurrence of the event ' A' and ' B' is $n_1 \times n_2$. Similarly each of the m_1 favourable cases can be associated with each of the m_2 favourable cases. So the total number of favourable cases for the event ' A' and ' B' is $m_1 \times m_2$

$$\therefore P(A \text{ f } B) = \frac{m_1 \; m_2}{n_1 \, n_2}$$

$$= \frac{m_1}{n_1} . \frac{m_2}{n_2}$$

$$= P(A).P(B)$$

Note:

The theorem can be extended to three or more independent events. If A,B,C....... be independent events, then P(A f B f C.......) = P(A).P(B).P(C).......

Note:

If A and B are independent then the complements of A and B are also independent. i.e $P(\overline{A} \cap \overline{B}) = P(\overline{A}) \,.\, P(\overline{B})$

1.5.2 Multiplication theorem for dependent events:

If A and B be two dependent events, i.e the occurrence of one event is affected by the occurrence of the other event, then the probability that both A and B will occur is

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

Proof:

Suppose an experiment results in n exhaustive, mutually exclusive and equally likely outcomes, m of them being favourable to the occurrence of the event A.

Out of these n outcomes let m_1 be favourable to the occurrence of another event B.

Then the outcomes favourable to the happening of the events ‘ A and B’ are m_1.

$$\therefore\ P(A \cap B) = \frac{m_1}{n}$$

$$= \frac{m_1}{n} \times \frac{m}{m} = \frac{m\,m_1}{n\,m}$$

$$= \frac{m}{n} \times \frac{m_1}{m}$$

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

Note:

In the case of three events A, B, C, $P(A \cap B \cap C) = P(A).\ P(B/A).\ P(C/A \cap B)$. ie., the probability of occurrence of A, B and C is equal to the probability of A times the probability of B given that A has occurred, times the probability of C given that both A and B have occurred.

1.6 BAYES’ Theorem:

The concept of conditional probability discussed earlier takes into account information about the occurrence of one event to predict the probability of another event. This concept can be extended to revise probabilities based on new information and to determine the probability that a particular effect was due to specific cause. The procedure for revising these probabilities is known as Bayes theorem.

The Principle was given by Thomas Bayes in 1763. By this principle, assuming certain prior probabilities, the posteriori probabilities are obtained. That is why Bayes' probabilities are also called posteriori probabilities.

Bayes' Theorem or Rule (Statement only):

Let A_1, A_2, A_3, ……A_i, ……A_n be a set of n mutually exclusive and collectively exhaustive events and $P(A_1)$, $P(A_2)$…. $P(A_n)$ are their corresponding probabilities. If B is another event such that P(B) is not zero and the priori probabilities $P(B|A_i)$ i =1,2….n are also known. Then

$$P(A_i \mid B) = \frac{P(B \mid A_i)\ P(A_i)}{\sum_{i=1}^{k} P(B \mid A_i)\ P(A_i)}$$

1.7 Basic principles of Permutation and Combination:

Factorial:

The consecutive product of first *n* natural numbers is known as ***factorial*** **n** and is denoted as **n!** or $\angle n$

That is $n! = 1 \times 2 \times 3 \times 4 \times 5 \times ... \times n$

$3! = 3 \times 2 \times 1$

$4! = 4 \times 3 \times 2 \times 1$

$5! = 5 \times 4 \times 3 \times 2 \times 1$

Also $5! = 5 \times (4 \times 3 \times 2 \times 1) = 5 \times (4!)$

Therefore this can be algebraically written as $n! = n \times (n - 1)!$

Note that $1! = 1$ and $0! = 1$.

Permutations:

Permutation means arrangement of things in different ways. Out of three things A, B, C taking two at a time, we can arrange them in the following manner.

A B B A

A C	C A
B C	C B

Here we find 6 arrangements. In these arrangements order of arrangement is considered. The arrangement AB and the other arrangement BA are different.

The number of arrangements of the above is given as the number of permutations of 3 things taken 2 at a time which gives the value 6. This is written symbolically, $_3P_2 = 6$

Thus the number of arrangements that can be made out of *n* things taken *r* at a time is known as the number of permutation of *n* things taken *r* at a time and is denoted as nPr.

The expansion of nPr is given below:

$nPr = n(n-1)(n-2)\ldots\ldots\ldots\ldots[n-(r-1)]$

The same can be written in factorial notation as follows:

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

For example, to find $_{10}P_3$ we write this as follows:

$$\begin{aligned} {}_{10}P_3 &= 10(10-1)(10-2) \\ &= 10 \times 9 \times 8 \\ &= 720 \end{aligned}$$

[To find $_{10}P_3$, Start with 10, write the product of 3 consecutive natural numbers in the descending order]

Simplifying $_{10}P_3$ using factorial notation:

$$_{10}P_3 = \frac{10!}{(10-3)!} = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{7\times6\times5\times4\times3\times2\times1}$$

$$= 10 \times 9 \times 8$$

$$= 720$$

Note that $nPo = 1$, $nP_1 = n$, $nP_n = n!$

Combinations:

A combination is a selection of objects without considering the order of arrangements.

For example, out of three things A,B,C we have to select two things at a time.

This can be selected in three different ways as follows:
A B A C B C

Here the selection of the object A B and B A are one and the same. Hence the order of arrangement is not considered in combination. Here the number of combinations from 3 different things taken 2 at a time is 3.

This is written symbolically $_3C_2 = 3$
Thus the number of combination of n different things, taken r at a time is given by $nCr = \dfrac{n\,Pr}{r!}$

Or $nCr = \dfrac{n!}{(n-r)!r!}$

Note that $nC_0 = 1$, $nC_1 = n$, $nC_n = 1$

Find $_{10}C_3$. $_{10}C_3 = \dfrac{_{10}P_3}{3!} = \dfrac{10\times9\times8}{1\times2\times3} = 120$

Find $_8C_4$. $_8C_4 = \dfrac{8\times7\times6\times5}{1\times2\times3\times4} = 70$

[To find $_8C_4$**:** In the numerator, first write the product of 4 natural numbers starting with 8 in descending order and in the denominator write the factorial 4 and then simplify.]
Compare $_{10}C_8$ and $_{10}C_2$

$$_{10}C_8 = \frac{10\times9\times8\times7\times6\times5\times4\times3}{1\times2\times3\times4\times5\times6\times7\times8} = \frac{10\times9}{1\times2} = 45$$

$$_{10}C_2 = \frac{10\times9}{1\times2} = 45$$

From the above, we find $_{10}C_8 = {_{10}C_2}$

This can be got by the following method also:
${}_{10}C_{8} = {}_{10}C_{(10-8)} = {}_{10}C_{2}$

This method is very useful, when the difference between *n* and *r* is very high in *n*C*r*.
This property of the combination is written as ${}_{n}C_{r} = {}_{n}C_{(n-r)}$.
To find ${}_{200}C_{198}$ we can use the above formula as follows:

$${}_{200}C_{198} = {}_{200}C_{(200-198)} = {}_{200}C_{2} = \frac{200\times199}{1\times2} = 19900.$$

Example:

Out of 13 players, 11 players are to be selected for a cricket team. In how many ways can this be done?
Out of 13 players, 11 players are selected in ${}_{13}C_{11}$ ways

$$\text{i.e.} \quad {}_{13}C_{11} = {}_{13}C_{2} = \frac{13 \times 12}{1 \times 2} = 78.$$

Example 1:

Three coins are tossed simultaneously Find the probability that
(i) no head (ii) one head (iii) two heads
(iv) atleast two heads. (v) atmost two heads appear.

Solution:

The sample space for the 3 coins is
S = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} ; n(S) = 8

(i) No head appear A = {TTT}; n(A) = 1

$$\therefore P(A) = \frac{1}{8}$$

(ii) One head appear B = {HTT, THT, TTH}; n (B) = 3

$$\therefore P(B) = \frac{3}{8}$$

(iii) Two heads appear C = {HHT, HTH, THH}; n(c)=3

$$\therefore P(C) = \frac{3}{8}$$

(iv) Atleast two heads appear
D = { HHT, HTH, THH, HHH}; n(D) = 4

$$\therefore\ P(D) = \frac{4}{8} = 1/2$$

(v) Atmost two heads appear E = { TTT, HTT, THT, TTH,HHT, HTH,THH}

n(E)= 7

$$\therefore P(E) = \frac{7}{8}$$

Example 2:

When two dice are thrown, find the probability of getting doublets (Same number on both dice)

Solution:

When two dice are thrown, the number of points in the sample space is n(S) = 36

Getting doublets: A = {(1,1) , (2,2) , (3,3) , (4,4) , (5,5) , (6,6)}

$$\therefore P(A) = \frac{6}{36} = \frac{1}{6}$$

Example 3:

A card is drawn at random from a well shuffled pack of 52 cards. What is the probability that it is (i) an ace (ii) a diamond card

Solution:

We know that the Pack contains 52 cards $\therefore$n(S)= 52

(i) There are 4 aces in a pack. n(A) = 4

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

(ii) There are 13 diamonds in a pack $\therefore$n(B) = 13

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

Example 4:

A ball is drawn at random from a box containing 5 green, 6 red, and 4 yellow balls. Determine the probability that the ball drawn is (i) green (ii) Red (iii) yellow (iv) Green or Red (v) not yellow.

Solution:

Total number of balls in the box = 5+6+4 = 15 balls

(i) Probability of drawing a green ball $= \dfrac{5}{15} = \dfrac{1}{3}$

(ii) Probability of drawing a red ball $= \dfrac{6}{15} = \dfrac{2}{5}$

(iii) Probability of drawing a yellow ball = $\dfrac{4}{15}$

(iv) Probability of drawing a Green or a Red ball $= \dfrac{5}{15} + \dfrac{6}{15} = \dfrac{11}{15}$

(v) Probability of getting not yellow = 1 – P (yellow)

$$= 1 - \frac{4}{15}$$

$$= \frac{11}{15}$$

Example 5:

Two dice are thrown, what is the probability of getting the sum being 8 or the sum being 10?

Solution:

Number of sample points in throwing two dice at a time is n(S)=36

Let A= {the sum being 8}

$\therefore$ A= {(6,2), (5,3) , (4,4), (3,5) , (2,6)}; P(A) = $\dfrac{5}{36}$

B = { the sum being 10}

$\therefore$ B = {(6,4), (5,5) (4,6)} ; P(B) = $\dfrac{3}{36}$

A f B = { 0 } ; n(A f B) = 0

$\therefore$ The two events are mutually exclusive

$\therefore$P(AUB) = P(A) + P(B)

$$= \frac{5}{36} + \frac{3}{36}$$

$$= \frac{8}{36} = \frac{2}{9}$$

Example 6 :

Two dice are thrown simultaneously. Find the probability that the sum being 6 or same number on both dice.

Solution:

n(S) = 36

The total is 6:

$\therefore$ A = {(5,1) , (4,2), (3,3) , (2,4) , (1,5)}; P(A) = $\frac{5}{36}$

Same number on both dice:

$\therefore$B = {(1,1) (2,2), (3,3), (4,4), (5,5), (6,6)}; P(B) = $\frac{6}{36}$

A f B = {(3,3)} ; P(A   B) =$\frac{1}{36}$

Here the events are not mutually exclusive.

$\therefore$ P(AUB) = P(A) + P(B) – P(A f B)

$$= \frac{5}{36} + \frac{6}{36} - \frac{1}{36}$$

$$= \frac{5+6-1}{36}$$

$$= \frac{11-1}{36}$$

$$= \frac{10}{36}$$

$$= \frac{5}{18}$$

Example 7:

Two persons A and B appeared for an interview for a job. The probability of selection of A is 1/3 and that of B is 1/2. Find the probability that

(i) both of them will be selected
(ii) only one of them will be selected
(iii) none of them will be selected

Solution:

$P(A)=\frac{1}{3}$, $P(B)=\frac{1}{2}$

$P(\overline{A})=\frac{2}{3}$ and $P(\overline{B})=\frac{1}{2}$

Selection or non-selection of any one of the candidate is not affecting the selection of the other. Therefore A and B are independent events.

(i) Probability of selecting both A and B

$$P(A \cap B) = P(A).P(B)$$
$$= \frac{1}{3}\times\frac{1}{2}$$
$$= \frac{1}{6}$$

(ii) Probability of selecting any one of them

= P (selecting A and not selecting B) + P(not selecting A and selecting B)

i.e $P(A \cap \overline{B}) + P(\overline{A} \cap B) = P(A).\ P(\overline{B}) + P(\overline{A}).\ P(B)$

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

(iii) Probability of not selecting both A and B

$$\text{i.e } P(\overline{A} \cap \overline{B})$$
$$= P(\overline{A}).\ P(\overline{B})$$
$$= \frac{2}{3}\cdot\frac{1}{2}$$
$$= \frac{1}{3}$$

Example 8:

There are three T.V programmes A , B and C which can be received in a city of 2000 families. The following information is available on the basis of survey.

1200 families listen to Programme A
1100 families listen to Programme B
800 families listen to Programme C
765 families listen to Programme A and B
450 families listen to Programme A and C
400 families listen to Programme B and C
100 families listen to Programme A, B and C

Find the probability that a family selected at random listens atleast one or more T.V Programmes.

Solution:

Total number of families n(S)= 2000

Let n(A) = 1200

n(B) = 1100

n(C) = 800

n(A f B) = 765

n(A f C) = 450

n(B f C) = 400

n(A f B f C) = 100

Let us first find n(AUBUC).

n(AUBUC) = n(A) + n(B)+ n(C) – n(A f B)–n(A f C)– n(B f C) + n(A f B f C)

= 1200 + 1100 + 800 – 765 – 450 – 400 + 100

n(AUBUC) = 1585

$$\text{now } P(AUBUC) = \frac{n(AUBUC)}{n(S)}$$

$$= \frac{1585}{2000} = 0.792$$

Therefore about 79% chance that a family selected at random listens to one or more T.V. Programmes.

Example 9:

A stockist has 20 items in a lot. Out of which 12 are non-defective and 8 are defective. A customer selects 3 items from the lot. What is the probability that out of these three items (i) three items are non-defective (ii) two are non defective and one is defective

Solution:

(i) Let the event, that all the three items are non-defective, be denoted by E_1. There are 12 non-defective items and out of them 3 can be selected in $12C_3$ ways ie $n(E_1)=12C_3$

Total number of ways in which 3 items can be selected are $20C_3$

i.e $n(S) = 20C_3$

$$\therefore P(E_1) = \frac{n(E_1)}{n(S)} = \frac{12C_3}{20C_3}$$

$$= \frac{12\times11\times10}{20\times19\times18}$$

$$= 0.193$$

ii) Let the event, that two items are non-defective and one is defective be denoted by E_2.

Two non-defective items out of 12 can be selected in$12C_2$ ways. One item out of 8 defective can be selected in $8C_1$ ways.

Thus $n(E_2) =12C_2 . 8C_1$

Then the probability $P(E_2) = \frac{n(E_2)}{n(S)} = \frac{12C_2 . 8C_1}{20C_3}$

$$= \frac{12\times11\times8\times3}{20\times19\times18}$$

$$= 0.463$$

Example 10:

A test paper containing 10 problems is given to three students A,B,C. It is considered that student A can solve 60% problems, student B can solve 40% problems and student C can solve 30% problems. Find the probability that the problem chosen from the test paper will be solved by all the three students.

Solution:

Probability of solving the problem by A = 60%
Probability of solving the problem by B = 40%
Probability of solving the problem by C = 30%

Solving the problem by a student is independent of solving the problem by the other students.

Hence, P(A f B f C) = P(A). P(B). P(C)

$$= \frac{60}{100} \times \frac{40}{100} \times \frac{30}{100}$$

$$= 0.6 \times 0.4 \times 0 .3$$

$$= 0.072$$

Example 11:

From a pack of 52 cards, 2cards are drawn at random. Find the probability that one is king and the other is queen.

Solution:

From a pack of 52 cards 2 cards are drawn n(S)=$52C_2$

Selection of one king is in $4C_1$ ways

Selection of one queen is in $4C_1$ ways

Selection of one king and one queen is in $4C_1.4C_1$ ways

ie n(E) = $4C_1.4C_1$

$$\therefore \text{ P(E)} = \frac{\text{n(E)}}{\text{n(S)}} = \frac{4C_1.4C_1}{52C_2}$$

$$= 4 \times 4 \div \frac{52 \times 51}{1 \times 2}$$

$$= \frac{4 \times 4 \times 2}{52 \times 51}$$

$$= \frac{8}{663}$$

Example 12:

An urn contains 4 black balls and 6 white balls. If 3 balls are drawn at random, find the probability that (i) all are black (ii) all are white

Solution:

Total number of balls = 10

Total number ways of selecting 3 balls = $10C_3$

(i) Number of ways of drawing 3 black balls = $4C_3$

Probability of drawing 3 black balls = $\frac{4C_3}{10C_3}$

$$= \frac{4\times3\times2}{1\times2\times3} \div \frac{10\times9\times8}{1\times2\times3}$$

$$= \frac{4\times3\times2}{10\times9\times8}$$

$$= \frac{1}{30}$$

(ii) Number of ways of drawing 3 white balls = $6C_3$

Probability of drawing 3 white balls $= \frac{6C_3}{10C_3}$

$$= \frac{6\times5\times4}{10\times9\times8}$$

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

Example 13:

A box containing 5 green balls and 3 red colour balls. Find the probability of selecting 3 green colour balls one by one (i) without replacement (ii) with replacement

Solution:

(i) Selection without replacement

Selecting 3 balls out of 8 balls = $8C_3$ ways

i.e n(S) = $8C_3$

Selecting 3 green balls in $5C_3$ ways

$\therefore$ P(3 green balls) = $\frac{5C_3}{8C_3} = \frac{5\times4\times3}{8\times7\times6} = \frac{5}{28}$

(ii) Selection with replacement

When a ball is drawn and replaced before the next draw, the number of balls in the box remains the same. Also the 3 events of drawing a green ball in each case is independent. $\therefore$Probability of drawing a green ball in each case is $\frac{5}{8}$

The event of selecting a green ball in the first, second and third event are same,

$\therefore$ Probability of drawing

$$3 \text{ green balls } = \frac{5}{8} \times \frac{5}{8} \times \frac{5}{8} = \frac{125}{512}$$

Example 14:

A box contains 5 red and 4 white marbles. Two marbles are drawn successively from the box without replacement and it is noted that the second one is white. What is the probability that the first is also white?

Solution:

If w_1, w_2 are the events ‘ white on the first draw’ , ‘ white on the second draw’ respectively.

Now we are looking for $P(w_1/w_2)$

$$P(w_1/w_2) = \frac{P(w_1 \text{ f } w_2)}{P(w_2)} = \frac{P(w_1).P(w_2)}{P(w_2)}$$

$$= \frac{(4/9)(3/8)}{(3/8)}$$

$$= \frac{4}{9}$$

Example 15:

A bag contains 6 red and 8 black balls. Another bag contains 7 red and 10 black balls. A bag is selected and a ball is drawn. Find the probability that it is a red ball.

Solution:

There are two bags

$\therefore$ probability of selecting a bag = $\frac{1}{2}$

Let A denote the first bag and B denote the second bag.

Then P(A) = P(B)= $\frac{1}{2}$

Bag ‘ A’ contains 6 red and 8 black balls.

$\therefore$ Probability of drawing a red ball is $\frac{6}{14}$

Probability of selecting bag A and drawing a red ball from that bag is P(A). P(R/A) = $\frac{1}{2} \times \frac{6}{14} = \frac{3}{14}$

Similarly probability of selecting bag B and drawing a red ball from that bag is P(B). P(R/B) = $\frac{1}{2} \times \frac{7}{17} = \frac{7}{34}$

All these are mutually exclusive events

∴ Probability of drawing a red ball either from the bag A or B is

P(R) = P(A) P(R/A) + P(B) P(R/B)

$$= \frac{3}{14} + \frac{7}{34}$$

$$= \frac{17\times 3 + 7\times 7}{238}$$

$$= \frac{51+49}{238}$$

$$= \frac{100}{238} = \frac{50}{119}$$

Example 16:

If P(A f B) = 0.3, P(A) = 0.6, P(B) = 0.7 Find the value of P(B/A) and P(A/B)

Solution:

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

$$= \frac{0.3}{0.6}$$

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

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

$$= \frac{0.3}{0.7}$$

$$= \frac{3}{7}$$

Example 17:

In a certain town, males and females form 50 percent of the population. It is known that 20 percent of the males and 5 percent of the females are unemployed. A research student studying the employment situation selects unemployed persons at random. What is the probability that the person selected is (i) a male (ii) a female?

Solution:

Out of 50% of the population 20% of the males are unemployed. i.e $\frac{50}{100} \times \frac{20}{100} = \frac{10}{100} = 0.10$

Out of 50% the population 5% of the females are unemployed.

$$\text{i.e } \frac{50}{100} \times \frac{5}{100} = \frac{25}{1000} = 0.025$$

Based on the above data we can form the table as follows:

	Employed	Unemployed	Total
Males	0.400	0.100	0.50
Females	0.475	0.025	0.50
Total	0.875	0.125	1.00

Let a male chosen be denoted by M and a female chosen be denoted by F

Let U denotes the number of unemployed persons then

(i) $P(M/U) = \frac{P(M \cap U)}{P(U)} = \frac{0.10}{0.125} = 0.80$

(ii) $P(F/U) = \frac{P(F \cap U)}{P(U)} = \frac{0.025}{0.125} = 0.20$

Example 18:

Two sets of candidates are competing for the positions on the Board of directors of a company. The probabilities that the first and second sets will win are 0.6 and 0.4 respectively. If the first set wins, the probability of introducing a new product is 0.8, and the corresponding probability if the second set wins is 0.3. What is the probability that the new product will be introduced?

Solution:

Let the probabilities of the possible events be:

$P(A_1)$ = Probability that the first set wins = 0.6

$P(A_2)$ = Probability that the second set wins = 0.4

$P(B)$ = Probability that a new product is introduced

$P(B/A_1)$ = Probability that a new product is introduced given that the first set wins = 0.8

$P(B/A_2)$ = Probability that a new product is introduced given that the second set wins = 0.3

Then the rule of addition gives:

P(new product) = P(first set and new product) + P(second set and new product)

$$\begin{aligned} \text{i.e } P(B) &= P(A_1 \;\; B) + P(A_2 \;\; B) \\ &= P(A_1)\, P(B/A_1) + P(A_2).P(B/A_2) \\ &= 0.6 \times 0.8 + 0.4 \times 0.3 \\ &= 0.60 \end{aligned}$$

Example 19:

Three persons A, B and C are being considered for the appointment as the chairman for a company whose chance of being selected for the post are in the proportion 4:2:3 respectively. The probability that A, if selected will introduce democratization in the company structure is 0.3 the corresponding probabilities for B and C doing the same are respectively 0.5 and 0.8. What is the probability that democratization would be introduced in the company?

Solution:

Let A_1 and A_2 and A_3 denote the events that the persons A, B and C respectively are selected as chairman and let E be the event of introducing democratization in the company structure.

Then we are given

$$P(A_1) = \frac{4}{9} \qquad P(A_2) = \frac{2}{9} \qquad P(A_3) = \frac{3}{9}$$

$$P(E/A_1) = 0.3 \qquad P(E/A_2) = 0.5 \qquad P(E/A_3) = 0.8$$

The event E can materialize in the following mutually exclusive ways:

(i) Person A is selected and democratization is introduced ie A_1 f E happens

(ii) Person B is selected and democratization is introduced ieA_2 f E happens

(iii) Person C is selected and democratization is introduced ie A_3 f E happens

Thus E = (A_1 f E) U (A_2 f E) U(A_3 f E) , where these sets are disjoint

Hence by addition rule of probability we have

P(E) = P(A_1 f E) + P(A_2 f E) + P(A_3 f E)

$$= P(A_1)\,P(E/A_1) + P(A_2)\,P(E/A_2) + P(A_3)\,P(E/A_3)$$

$$= \frac{4}{9} \times 0.3 + \frac{2}{9} \times 0.5 + \frac{3}{9} \times 0.8$$

$$= \frac{46}{90} = \frac{23}{45}$$

Example 20:

In a bolt factory machines $A_{1,}$ A_2, A_3 manufacture respectively 25%, 35% and 40% of the total output. Of these 5, 4, and 2 percent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What is the probability that it was manufactured by machine A_2 ?

Solution:

$P(A_1)$ = P(that the machine A_1 manufacture the bolts) = 25% = 0.25

Similarly $P(A_2) = 35\% = 0.35$ and

$P(A_3) = 40\% = 0.40$

Let B be the event that the drawn bolt is defective.

$P(B/A_1)$ = P (that the defective bolt from the machine A_1)

= 5 % = 0.05

Similarly, $P(B/A_2) = 4\% = 0.04$

And $P(B/A_3) = 2\% = 0.02$

We have to find $P(A_2/ B)$.

Hence by Bayes' theorem, we get

$$P(A_2/ B).=\frac{P(A_2)P(B/A_2)}{P(A_1)P(B/A_1)+P(A_2)P(B/A_2)+P(A_3)P(B/A_3)}$$

$$= \frac{(0.35)(0.04)}{(0.25)(0.05)+(0.35)(0.04)+(0.4)(0.02)}$$

$$= \frac{28}{69}$$

$$= 0.4058$$

Example 21:

A company has two plants to manufacture motorbikes. Plant I manufactures 80 percent of motor bikes, and plant II manufactures 20 percent. At Plant I 85 out of 100 motorbikes are rated standard quality or better.

At plant II only 65 out of 100 motorbikes are rated standard quality or better.

(i) What is the probability that the motorbike, selected at random came from plant I. if it is known that the motorbike is of standard quality?

(ii) What is the probability that the motorbike came from plant II if it is known that the motor bike is of standard quality?

Solution:

Let A_1 be the event of drawing a motorbike produced by plant I.

A_2 be the event of drawing a motorbike produced by plant II.

B be the event of drawing a standard quality motorbike produced by plant I or plant II.

Then from the first information, $P(A_1) = 0.80$, $P(A_2) = 0.20$

From the additional information

$P(B/A_1) = 0.85$

$P(B/A_2) = 0.65$

The required values are computed in the following table.

The final answer is shown in last column of the table.

Event	Prior probability P(A_i)	Conditional probability of event B given A_i P(B/A_i)	Joint probability P(A_i f B) = P(A_i)P(B/A_i)	Posterior (revised) probability P(A_i/B) = $\frac{P(A \cap B)}{P(B)}$
A_1	0.80	0.85	0.68	$\frac{0.68}{0.81} = \frac{68}{81}$
A_2	0.20	0.65	0.13	$\frac{0.13}{0.81} = \frac{13}{81}$
Total	1.00		P(B) = 0.81	1

Without the additional information, we may be inclined to say that the standard motor bike is drawn from plant I output, since P(A_1) = 80% is larger than P(A_2) =20%

Remark:

The above answer may be verified by actual number of motorbikes as follows:

Suppose 10,000 motorbikes were produced by the two plants in a given period, the number of motorbikes produced by plant I is

$10{,}000 \times 80\% = 8000$

and number of motorbikes produced by plant II is

$10000 \times 20\% = 2000 = 2000$

The number of standard quality motorbikes produced by plant I is

$$8000 \times \frac{85}{100} = 6800$$

And by plant II is

$$2000 \times \frac{65}{100} = 1300$$

The probability that a standard quality motor bike was produced by plant I is

$$\frac{6800}{6800+1300} = \frac{6800}{8100} = \frac{68}{81}$$

And that by plant II is

$$\frac{1300}{6800+1300} = \frac{1300}{8100} = \frac{13}{81}$$

The process of revising a set of prior probabilities may be repeated if more information can be obtained. Thus Bayes' theorem provides a powerful method in improving quality of probability for aiding the management in decision making under uncertainty.

Exercise - 1

I. Choose the best answer :

1. Probability is expressed as
   (a) ratio (b) percentage
   (c) Proportion (d) all the above
2. Probability can take values from
   (a) - ∞ to +∞ (b) - ∞ to 1
   (c) 0 to +1 (d) –1 to +1
3. Two events are said to be independent if
   (a) each out come has equal chance of occurrence
   (b) there is the common point in between them
   (c) one does not affect the occurrence of the other.
   (d) both events have only one point
4. Classical probability is also known as
   (a) Statistical probability
   (b) A priori probability
   (c) Empirical probability
   (d) None of the above
5. When a coin and a die are thrown, the number of all possible cases is
   (a) 7 (b) 8 (c)12 (d) 0

6. Probability of drawing a spade queen from a well shuffled pack of cards is

(a) $\frac{1}{13}$ (b) $\frac{1}{52}$ (c) $\frac{4}{13}$ (d) 1

7. Three dice are thrown simultaneously the probability that sum being 3 is

(a) 0 (b) 1 / 216 (c) 2 / 216 (d) 3 / 216

8. An integer is chosen from 1 to 20. The probability that the number is divisible by 4 is

(a) $\frac{1}{4}$ (b) $\frac{1}{3}$ (c) $\frac{1}{2}$ (d) $\frac{1}{10}$

9. The conditional probability of B given A is

(a) $\frac{P(A \cap B)}{P(B)}$ (b) $\frac{P(A \cap B)}{P(A)}$

(c) $\frac{P(AUB)}{P(B)}$ (d) $\frac{P(AUB)}{P(A)}$

10. P(X) = 0.15, P(Y) = 0.25, P(X f Y) = 0.10 then P(XUY) is

(a) 0.10 (b) 0.20 (c) 0.30 (d) 0.40

11. If P(A) = 0.5, P(B) = 0.3 and the events A and B are independent then P(A f B) is

(a) 0.8 (b) 0.15 (c) 0.08 (d) 0.015

12. If P(A) = 0.4 P(B) = 0.5 and P(A f B) = 0.2 then P(B/A) is

(a) $\frac{1}{2}$ (b) $\frac{1}{3}$ (c) $\frac{4}{5}$ (d) $\frac{2}{5}$

13. A coin is tossed 6 times. Find the number of points in the sample space.

(a) 12 (b)16 (c) 32 (d) 64

14. When a single die is thrown the event of getting odd number or even number are

(a) Mutually exclusive events
(b) Not-mutually exclusive events
(c) Independent event
(d) Dependent event

15. The probability of not getting 2, when a die is thrown is

(a) $\frac{1}{3}$ (b) $\frac{2}{3}$ (c) $\frac{1}{6}$ (d) $\frac{5}{6}$

II. Fill in the blanks:

16. The probability of a sure event is __________
17. The probability of an impossible event is _________
18. Mathematical probability is also called a _______________ probability.
19. The joint occurrence of two or more events is called ___________
20. If A and B are mutually exclusive events, then P(AUB) = ____
21. If A and B are independent events then P(A f B) = _________
22. If A and B are dependent events then P(A/B) = ___________
23. If A and B are mutually exclusive events P(A f B) = _________
24. When three coins are tossed the probability of getting 3 heads is _________
25. When three dice are thrown the probability of sum being 17 is __________
26. The probability getting the total is 11 when two dice are throws ____________

III Answer the following:

27. Define the following terms:

    Event, equally likely events, mutually exclusive events, exhaustive events, sample space.
28. Define dependent and independent events.
29. Define mathematical probability.
30. Define statistical probability.
31. State the axioms of probability.
32. Explain addition theorem on probability for any two events.
33. State the multiplication theorem on probability.
34. Define conditional probability.
35. State Bayes' Rule.

36. There are 5 items defective in a sample of 30 items. Find the probability that an item chosen at random from the sample is (i) defective, (ii) non-defective

37. Four coins are tossed simultaneously what is the probability of getting (i) 2 heads (ii) 3 heads (iii) atleast 3 heads

38. Two dice are thrown. What is probability of getting (i) the sum is 10 (ii) atleast 10

39. Three dice are rolled once. What is the chance that the sum of the face numbers on the dice is (i) exactly 18 (ii) exactly 17 (iii) atmost 17.

40. An integer is chosen from 20 to 30. Find the probability that it is a prime number.

41. An integer is chosen from 1 to 50. Find the probability that it is multiple of 5 or multiply of 7

42. From a pack of cards find the probability of drawing a spade card or a diamond card.

43. Find the probability that a leap year selected at random will contain 53 Sundays.

44. Find the probability that a non-leap year selected at random will contain either 53 Sundays or 53 Mondays.

45. If two events A and B are not mutually exclusive and are not connected with one random experiment P(A)= 1/4, P(B) =2/5 and P(AUB) = 1/2 then find the value of P(B/A)

46. For two independent events A and B for which P(A) = 1/2 and P(B) = 1/3. Find the probability that one of them occur.

47. For two events A and B, P(A) =1/3 = P($\overline{B}$), P(B/A) =1/4 find P(A/B)

48. A box contains 4 red pens and 5 black pens. Find the probability of drawing 3 black pens one by one (i) with replacement (ii) without replacement

49. An urn contains 5 red and 7 green balls. Another urn contains 6 red and 9 green balls. If a ball is drawn from any one of the two urns, find the probability that the ball drawn is green.

50. Two cards are drawn at random from a pack of 52 cards. Find the probability that the cards drawn are (i) a diamond and a spade (ii) a king and a queen (iii) 2 aces

51. A problem in statistics is given to two students A and B. The probability that A solves the problem is 1/2 and that of B' s to solve it is 2/3. Find the probability that the problem is solved.

52. A bag contains 6 white, 4 green and 10 yellow balls. Two balls are drawn at random. Find the probability that both will be yellow.

53. In a certain class there are 21 students in subject A, 17 in subject B and 10 in subject C. Of these 12 attend subjects A and B, 5 attend subjects B and C, 6 attend subjects A and C. These include 2 students who attend all the three subjects. Find the probability that a student studies one subject alone.

54. If P(A) = 0.3, P(B) = 0.2 and P(C) = 0.1 and A,B,C are independent events, find the probability of occurrence of at least one of the three events A , B and C

55. The odds that A speaks the truth are 3:2 and the odds that B speaks the truth 5:3. In what percentage of cases are they likely to contradict each other on an identical point?

56. The chances of X, Y and Z becoming managers of a certain company are 4:2:3. The probabilities that bonus scheme will be introduced if X, Y and Z become managers are 0.3, 0.5 and 0.4 respectively. If the bonus scheme has been introduced what is the probability that Z is appointed as the manager?

57. A manufacturing firm produces steel pipes in three plants with daily production volumes of 500, 1000 and 2000 units respectively. According to past experience, it is known that the fractions of defective outputs produced by the three plants are respectively, 0.005, 0.008 and 0.010. If a pipe is selected from days total production and found to be defective, what is the probability that it came from the (i) first plant (ii) the second plant (iii) the third plant?

Answers:

I.

1. (d)   2. (c)   3. (c)   4. (b)   5. (c)

6. (b)   7. (b)   8. (a)   9. (b)   10. (c)

11. (b)   12.(a)   13.(d)   14.(a)   15. (d)

II.

16. 1   17. zero   18. priori   19. Compound events

20. P(A) +P(B)   21. P(A). P(B)

22. $\frac{P(A \cap B)}{P(B)}$ 23. 0   24. $\frac{1}{8}$

25. $\frac{3}{216}$ 26. $\frac{1}{18}$

III

36. $\frac{1}{6}, \frac{5}{6}$ 37. $\frac{3}{8};\frac{1}{4};\frac{5}{16}$ 38. $\frac{1}{12}; \frac{1}{6}$ 39. $\frac{1}{216},\frac{3}{216},\frac{215}{216}$

40. $\frac{2}{11}$ 41. $\frac{8}{25}$ 42. $\frac{1}{2}$ 43. $\frac{2}{7}$ 44. $\frac{2}{7}$

45. P(A f B) = 3/20 ; P(B/A) = 3/5

46. $\frac{1}{2}$ 47. P(A f B) = 1/12   P(A/B)=1/8

48. $\frac{125}{729},\frac{5}{42}$ 49. $\frac{71}{120}$ 50. $\frac{13}{102};\frac{8}{663};\frac{1}{221}$

51. $\frac{5}{6}$ 52. $\frac{9}{38}$ 53. $\frac{8}{27}$

54. 0.496   55. $\frac{3}{5}\times\frac{3}{8}+\frac{2}{5}\times\frac{5}{8}=\frac{19}{40}$ = 47.5%

56. 6/17   57. (a) $\frac{1}{7};\frac{2}{7};\frac{4}{7}$ (b) $\frac{5}{61};\frac{16}{61};\frac{40}{61}$

Activity:

We know that, when a coin is tossed, the probability of getting a head is 0.5 mathematically.

Now do the following experiment.

1. Toss a fair coin 10 times. Record the event that the number of heads occur in this experiment
2. Toss a fair coin 100 time with the help of your friends group and record the same event that the number of heads appearing.
3. Now compare all the three above mentioned and write your inference.

2. RANDOM VARIABLE AND MATHEMATICAL EXPECTATION

2.0 Introduction:

It has been a general notion that if an experiment is conducted under identical conditions, values so obtained would be similar. Observations are always taken about a factor or character under study, which can take different values and the factor or character is termed as variable.

These observations vary even though the experiment is conducted under identical conditions. Hence, we have a set of outcomes (sample points) of a random experiment. A rule that assigns a real number to each outcome (sample point) is called random variable.

From the above discussion, it is clear that there is a value for each outcome, which it takes with certain probability. Hence a list of values of a random variable together with their corresponding probabilities of occurrence, is termed as Probability distribution**.**

As a tradition, probability distribution is used to denote the probability mass or probability density, of either a discrete or a continuous variable.

The formal definition of random variable and certain operations on random variable are given in this chapter prior to the details of probability distributions.

2.1 Random variable:

A variable whose value is a number determined by the outcome of a random experiment is called a random variable.

We can also say that a random variable is a function defined over the sample space of an experiment and generally assumes different values with a definite probability associated with each value. Generally, a random variable is denoted by capital letters like X, Y, Z…., where as the values of the random variable are denoted by the corresponding small letters like x, y, z …….

Suppose that two coins are tossed so that the sample space is S = {HH, HT, TH, TT}
Suppose X represent the number of heads which can come up, with each sample point we can associate a number for X as shown in the table below:

Sample point	HH	HT	TH	TT
X	2	1	1	0

Thus the random variable X takes the values 0, 1,2 for this random experiment.

The above example takes only a finite number of values and for each random value we can associate a probability as shown in the table.

Usually, for each random variable $x_{i,}$ the probability of respective random variable is denoted by $p(x_i)$ or simply $p_{i\,.}$

X	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$
$p(x_i)$	$p(x_i) = \frac{1}{4}$	$p(x_i) = \frac{2}{4}$	$p(x_i) = \frac{1}{4}$

Observe that the sum of the probabilities of all the random variable is equal to one. ie $p(x_1) + p(x_2) + p(x_3) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1$

Thus the probability distribution for a random variable provides a probability for each possible value and that these probabilities must sum to 1.

Similarly if 3 coins are tossed, the random variable for getting head will be X=0, X=1, X=2, X=3 and sum of their respective probabilities i.e $\Sigma p(x_i) = 1$

If two dice are rolled then the sample space S consists of 36 sample points. Let X denote the sum of the numbers on the two dice. Then X is a function defined on S by the rule X(i,j) = i+j . Then X is a random variable which can takes the values 2,3,4.....12. That is the range of X is {2,3,4.....12}

2.1.1 Discrete random variable:

If a random variable takes only a finite or a countable number of values, it is called a discrete random variable.

For example, when 3 coins are tossed, the number of heads obtained is the random variable X assumes the values 0,1,2,3 which form a countable set. Such a variable is a discrete random variable.

2.1.2 Continuous random variable:

A random variable X which can take any value between certain interval is called a continuous random variable.

Note that the probability of any single value at x, value of X is zero. i.e $P(X = x) = 0$ Thus continuous random variable takes value only between two given limits.

For example the height of students in a particular class lies between 4 feet to 6 feet.

We write this as $X = \{x|4 \leq x \leq 6\}$

The maximum life of electric bulbs is 2000 hours. For this the continuous random variable will be $X = \{x \mid 0 \leq x \leq 2000\}$

2.2 Probability mass function:

Let X be a discrete random variable which assumes the values $x_1, x_2, \ldots x_n$ with each of these values, we associate a number called the probability $P_i= P(X=x_i)$, $i = 1,2,3\ldots n$ This is called probability of x_i satisfying the following conditions.

(i) $P_i \geq 0$ for all i, ie P_i' s are all non-negative

(ii) $\Sigma p_i = p_1 + p_2 + \ldots p_n =1$
ie the total probability is one.

This function p_i or $p(x_i)$ is called the **probability mass function** of the discrete random variable X.

The set of all possible ordered pairs (x, p(x)) is called the probability distribution of the random variable X.

Note:

The concept of probability distribution is similar to that of frequency distribution. Just as frequency distribution tells us how the total frequency is distributed among different values (or classes) of the variable, a probability distribution tells us how total probability 1 is distributed among the various values which the random variable can take. It is usually represented in a tabular form given below:

X	x_1	x_2	x_3	….	x_n
P(X = x)	$P(x_1)$	$P(x_2)$	$P(x_3)$	….	$P(x_n)$

2.2.1 Discrete probability distribution:

If a random variable is discrete in general, its distribution will also be discrete. For a discrete random variable X, the distribution function or cumulative distribution is given by F(x) and is written as $F(x) = P(X \leq x)$; $-\infty < x < \infty$

Thus in a discrete distribution function, there are a countable number of points $x_1, x_2, \ldots$ and their probabilities p_i such that

$$F(x_i) = \sum_{x_i < x} p_i \, , \quad i = 1, 2, \ldots\ldots n$$

Note:

For a discrete distribution function, $F(x_j) - F(x_{j-1}) = p(x_j)$

2.2.2 Probability density function (pdf):

A function *f* is said to be the probability density function of a continuous random variable X if it satisfies the following properties.

(i) $f(x) \geq 0 \quad -\infty < x < \infty$

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

Remark:

In case of a discrete random variable, the probability at a point ie P(x = *a*) is not zero for some fixed ' a' However in case of continuous random variables the probability at a point is always zero

ie $P(X = a) = \int_{a}^{a} f(x)\,dx = 0$

Hence $P(a \le X \le b) = P(a < X < b) = P(a \le X < b) = P(a < X \le b)$

The probability that x lies in the interval (a,b) is given by

$$P(a < X < b) = \int_a^b f(x)\, dx$$

Distribution function for continuous random variable.

If X is a continuous random variable with p.d.f f(x), then the distribution function is given by

(i) $F(x) = \int_{-\infty}^{x} f(x)\, dx = P(X \le x)$; $-\infty < x < \infty$

(ii) $F(b) - F(a) = \int_a^b f(x)\, dx = P(a \le X \le b)$

2.3 Properties of distribution function:

Suppose that X be a discrete or continuous random variable, then

(i) F(x) is a non - decreasing function of x

(ii) $0 \le F(x) \le 1$, $-\infty < x < \infty$

(iii) $F(-\infty) = \underset{x \to -\infty}{\text{limit}} F(x) = 0$

(iv) $F(\infty) = \underset{x \to \infty}{\text{limit}} F(x) = 1$

(v) If F(x) is the cumulative distribution function of a continuous random variable X with p.d.f f(x) then $F'(x) = f(x)$

Example 1:

A random variable has the following probability distribution

Values of X	0	1	2	3	4	5	6	7	8
P(x)	a	$3a$	$5a$	$7a$	$9a$	$11a$	$13a$	$15a$	$17a$

(1) Determine the value of a

(2) Find (i) $P(x < 3)$ (ii) $P(x \le 3)$ (iii) $P(x > 7)$
(iv) $P(2 \le x \le 5)$, (v) $P(2 < x < 5)$

(3) Find the cumulative distribution function of x.

Solution:

(1) Since p_i is the probability mass function of discrete random variable X,

We have $\Sigma p_i = 1$

$\therefore\ a + 3a + 5a + 7a + 9a + 11a + 13a + 15a + 17a = 1$

$81a = 1$

$a = 1/81$

(2)

(i) $P(x < 3) = P(x=0) + P(x=1) + P(x=2)$

$= a + 3a + 5a$

$= 9a$

$= 9\left(\frac{1}{81}\right)$

$= \frac{1}{9}$

(ii) $P(x \le 3) = P(x=0) + P(x=1) + P(x=2) + P(x=3)$

$= a + 3a + 5a + 7a$

$= 16a$

$= \frac{16}{81}$

iii) $P(x > 7) = P(x = 8)$

$= 17a$

$= \frac{17}{81}$

iv) $P(2 \le x \le 5) = P(x=2) + P(x=3) + P(x = 4) + P(x=5)$

$= 5a + 7a + 9a + 11a$

$= 32a$

$= \frac{32}{81}$

v) $P(2 < x < 5) = P(x = 3) + P(x = 4)$

$= 7a + 9a$

$= 16a$

$= \frac{16}{81}$

3) The distribution function is as follows:

X=x	0	1	2	3	4	5	6	7	8
F(x)= P (X≤ x)	a	$4a$	$9a$	$16a$	$25a$	$36a$	$49a$	$64a$	$81a$
(or) F(x)	$\frac{1}{81}$	$\frac{4}{81}$	$\frac{9}{81}$	$\frac{16}{81}$	$\frac{25}{81}$	$\frac{36}{81}$	$\frac{49}{81}$	$\frac{64}{81}$	$\frac{81}{81}=1$

Example 2:

Find the probability distribution of the number of sixes in throwing two dice once.

Solution:

When two dice are thrown the total number of sample points are 36.

Let X denote the number of sixes obtained in throwing two dice once. Then X is the random variable, which can take the values 0,1,2.

Let A denote the success of getting a six in throwing a die and $\overline{A}$ denote not getting a six.

Then probability getting a six

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

Probability not getting a six

$$P(\overline{A}) = \frac{5}{6}$$

No sixes:

$$\therefore P(x = 0) = P(\overline{A} \text{ and } \overline{A})$$

$$= P(\overline{A}) \,.\, P(\overline{A})$$

$$= \frac{5}{6} \,.\, \frac{5}{6}$$

$$= \frac{25}{36}$$

$$P(x = 1) = P(A \text{ and } \overline{A}) \text{ or } P(\overline{A} \text{ and } A)$$

$$= P(A)\,.\,P(\overline{A}) + P(\overline{A})\,.P(A)$$

$$= \frac{1}{6}.\ \frac{5}{6} + \frac{5}{6}\ \frac{1}{6}$$

$$= \frac{5}{36} + \frac{5}{36}$$

$$= \frac{10}{36}$$

$$= \frac{5}{18}$$

$$P(x = 2) = P(A \text{ and } A)$$

$$= P(A)\,.P(A)$$

$$= \frac{1}{6}.\frac{1}{6}$$

$$= \frac{1}{36}$$

Hence the probability distribution of X is given by

X= x	0	1	2
P(X = x)	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$

Example 3:

An urn contains 6 red and 4 white balls. Three balls are drawn at random. Obtain the probability distribution of the number of white balls drawn.

Solution:

The total number of balls in the urn is 10

Let X denote the number of white balls drawn

If three balls are drawn, the random variable takes the value X= 0, 1, 2, 3

Probability of getting white balls from the urn containing 10 balls (red and white) with the following combination are

$$P\text{ (no white, 3 red balls)} = \frac{4C_0 6C_3}{10C_3} = \frac{1\times120}{720} = \frac{5}{30}$$

$$P\text{ (1 white, 2 red)} = \frac{4C_1.6C_2}{10C_3} = \frac{15}{30}$$

$$P\text{ (2 white, 1 red)} = \frac{4C_2 6C_1}{10C_3} = \frac{9}{30}$$

$$P\text{ (3 white, no red)} = \frac{4C_3 6C_0}{10C_3} = \frac{1}{30}$$

Hence the probability distribution of X is given by

C	0	1	2	3
P(X=x)	$\frac{5}{30}$	$\frac{15}{30}$	$\frac{9}{30}$	$\frac{1}{30}$

2.4 An introduction to elementary calculus:

Before going to see the problems on continuous random variables, we need to know some fundamental ideas about differentiation and integration, which are part of calculus in higher-level mathematics.

Hence we introduce some simple techniques and formulae to calculate the problems in statistics, which involve calculus methods.

2.4.1 Differentiation:

1. Functional value is an exact value. For some function f(x), when x = *a*, we obtain the functional value as f(*a*) = k.
2. Limiting value is an approximate value. This value approaches the nearest to the exact value k.

   Suppose the exact value is 4. Then the limiting value will be 4.000000001 or 3.999999994. Here the functional value and limiting value are more or less same.

   Hence in many occasions we use the limiting values for critical problems.

   The limiting value of f(x) when x approaches a number 2 is given as

   $$\underset{x à 2}{\text{Limit}}\ f(x) = f(2) = l$$

3. The special type of existing limit, $\underset{h\to 0}{\text{limit}} \dfrac{f(x+h)-f(x)}{h}$ is called the derivative of the function f with respect to x and is denoted by f ′(x). If y is the function x then we say the differential coefficient of y with respect to x and is denoted as $\dfrac{dy}{dx}$

4. Some rules on differentiation:
   - (i) Derivative of a constant function is zero. f ′(c)=0 where c is some constant.
   - (ii) If u is a function of x and k is some constant and dash denotes the differentiation, [ku]′ = k[u]′
   - (iii) $(u \pm v)' = u' \pm v'$
   - (iv) $(uv)' = u'v + uv'$
   - (v) $\left[\dfrac{u}{v}\right]' = \dfrac{u'v - uv'}{v^2}$

5. Important formulae:
   - (i) $(x^n)' = nx^{n-1}$
   - (ii) $(e^x)' = e^x$
   - (iii) $(\log_x)' = \dfrac{1}{x}$

Example 4:

Evaluate the following limits:

(i) $\underset{x\to 2}{\text{Limit}} \dfrac{x^2+5x}{x+2}$ (ii) $\underset{x\to 1}{\text{Limit}} \dfrac{x^2-1}{x-1}$

Solution:

(i) $\underset{x\to 2}{\text{Limit}} \dfrac{x^2+5x}{x+2} = \dfrac{(2)^2+5(2)}{2+2} = \dfrac{4+10}{4} = \dfrac{14}{4} = \dfrac{7}{2}$

(ii) $\underset{x\to 1}{\text{Limit}} \dfrac{x^2-1}{x-1} = \dfrac{1^2-1}{1-1} = \dfrac{0}{0}$. This is an indeterminate form.

Therefore first factorise and simplify and then apply the same limit to get the limiting value

$$\therefore \quad \frac{x^2-1}{x-1} = \frac{(x-1)(x+1)}{(x-1)} = x+1$$

$$\therefore \ \underset{x\to 1}{\text{Limit}} \frac{x^2-1}{x-1} = \underset{x\to 1}{\text{Limit}}\,(x+1) = 1+1 = 2$$

Example 5:

Find the derivative of the following with respect to x.

(i) $x^{12} + 7$ (ii) $(x^4 + 4x^2 - 5)$ (iii) $(x^3)\,(e^x)$ (iv) $\frac{x^2+1}{x-5}$

Solution:

(i) Let $y = x^{12} + 7$

$$\therefore \ \frac{dy}{dx} = 12x^{12-1} + 0 = 12x^{11}$$

(ii) Let $y = x^3 + 4x^2 - 5$

$$y' = 4x^3 + 4(2x) - 0$$
$$= 4x^3 + 8$$

(iii) Let $y = x^3\, e^x$

$$(uv)' = u'v + uv'$$
$$= [x^3]'\,(e^x) + (x^3)\,[e^x]'$$
$$= 3x^2\, e^x + x^3\, e^x$$
$$= e^x\,(3x^2 + x^3)$$

(iv) $y = \frac{x^2+1}{x-5}$. This is of the type $\left[\frac{u}{v}\right]' = \frac{u'v - uv'}{v^2}$

$$\therefore \ \frac{dy}{dx} = \frac{\left[x^2+1\right]'(x-5) - (x^2+1)\left[x-5\right]'}{(x-5)^2}$$

$$= \frac{[2x](x-5) - (x^2+1)[1]}{(x-5)^2}$$

$$= \frac{2x^2 - 10x - x^2 - 1}{(x-5)^2}$$

$$= \frac{x^2 - 10x - 1}{(x-5)^2}$$

2.4.2 Integration:

Integration is known as the reverse process of differentiation. Suppose the derivative of x^3 is $3x^2$. Then the integration of $3x^2$ with respect to x is x^3 . We write this in symbol as follows:

$$\frac{d}{dx}(x^4) = 4x^3 \qquad \Rightarrow 4 \int x^3 dx = x^4$$

Similarly

$$\frac{d}{dx}(x^8) = 8x^7 \qquad \Rightarrow 8 \int x^7 dx = x^8$$

$$\frac{d}{dx}(e^x) = e^x \qquad \Rightarrow \int e^x dx = e^x$$

Note:

While differentiating the constant term we get zero. But in the reverse process, that is on integration, unless you know the value of the constant we cannot include. That is why we include an arbitrary constant C to each integral value.

Therefore the above examples, we usually write $\int e^x \, dx = e^x + c$ and $\int 8x^7 \, dx = x^8 + c$

These integrals are also called improper integrals or indefinite integrals

Rules and formulae on integration:

(i) $\int k \, dx = kx$

(ii) $\int x^n \, dx = \frac{x^{n+1}}{n+1}$

(iii) $\int e^x \, dx = e^x$

(iv) $\int \frac{1}{x} dx = \log x$

(v) $\int (u \pm v) dx = \int u \, dx \pm \int v \, dx$

Example 6:

Integrate the following with respect to x:

(i) $\int x^6 \, dx = \dfrac{x^{6+1}}{6+1} = \dfrac{x^7}{7} + c$

(ii) $\int x^{-5} \, dx = \dfrac{x^{-5+1}}{-5+1} = \dfrac{x^{-4}}{-4} = - \dfrac{1}{4}\dfrac{1}{x^4} = -\dfrac{1}{4x^4} + c$

(iii) $\int \dfrac{1}{x} dx = \log x + c$

(iv) $\int \sqrt{x} dx = \int x^{1/2} dx = \dfrac{x^{1/2+1}}{\frac{1}{2}+1} = \dfrac{x^{3/2}}{\frac{3}{2}} = \dfrac{2}{3} x^{3/2} + c$

(v) $\int (x^4 + 2x^2 + 4x + 8) \, dx$

$$= \frac{x^5}{5} + 2\frac{x^3}{3} + 4\frac{x^2}{2} + 8x + c$$

(vi) $\int (e^x + x^4 + 1/x^3 + 10) \, dx$

$= e^x + x^5/5 - 1/2x^2 + 10x + c$

The above discussed integrals are known as improper integrals or indefinite integrals. For the proper or definite integrals we have the limiting point at both sides. ie on the lower limit and the upper limit.

This integral $\int f(x) dx$ is an indefinite integral

Integrating the same function within the given limits *a* and *b* is known as the definite integral.

ie $\int_a^b f(x) dx$ = k (a constant value) is a definite integral where *a* is known as lower limit and *b* is known as the upper limit of the definite integral.

To find the value of definite integral we use the formulae:

Suppose $\int f(x) dx = F(x)$ then $\int_a^b f(x) dx = F(b) - F(a)$

An important note to the Teachers and students

As per as statistics problems concerned, the differentiation and integration methods restricted to simple algebraic functions only.

Example 7:

Evaluate the following definite integrals.

(i) $\int_0^4 3x^2 dx$ (ii) $\int_1^3 x^3 dx$ (iii) $\int_2^5 x dx$

Solution:

(i) $\int_0^4 3x^2 dx = \left[\frac{3x^3}{3}\right]_0^4 = [x^3]_0^4$

$= 4^3 - 0^3 = 64$

(ii) $\int_1^3 x^3 dx = \left[\frac{x^4}{4}\right]_1^3$

$= \frac{1}{4}[x^4]_1^3$

$= \frac{1}{4}[3^4 - 1^4]$

$= \frac{1}{4}[81-1]$

$= \frac{1}{4}[80]$

$= 20$

(iii) $\int_2^5 x dx = \left[\frac{x^2}{2}\right]_2^5$

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

$= \frac{1}{2}[25-4] = \frac{21}{2}$

Example 8:

Examine whether $f(x) = 5x^4$, $0 < x < 1$ can be a p.d.f of a continuous random variable x.

Solution:

For a probability density function, to show that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

That is to show that $\int_0^1 5(x)^4\,dx = 1$

$$\int_0^1 5(x)^4\,dx \quad = 5\left[\frac{x^5}{5}\right]_0^1$$

$$= \frac{5}{5}\left[x^5\right]_0^1$$

$$= [1^5 - 0]$$

$$= 1$$

$\therefore$ f(x) is a p.d.f

Example 9:

A continuous random variable x follows the rule $f(x) = Ax^2$, $0 < x < 1$. Determine A

Solution:

Since f(x) is a p.d.f, $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Therefore $\int_0^1 Ax^2\,dx \quad = 1$

$$A\left[\frac{x^3}{3}\right]_0^1 \quad = 1$$

$$\frac{A}{3}\left[x^3\right]_0^1 \quad = 1$$

$$\frac{A}{3}[1] \quad = \quad 1$$

$$A \ = 3$$

Example 10:

Let f(x) = c(1-x) x^2 , 0 < x < 1 be a probability density function of a random variable x. Find the constant c

Solution:

$f(x) = c(1-x)x^2 , 0 < x < 1$

since f(x) is a p.d.f $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\therefore \quad \int_0^1 c(x^2 - x^3)dx = 1$$

$$c\left(\frac{x^3}{3} - \frac{x^4}{4}\right)_0^1 = 1$$

$$c\left[\left(\frac{1^3}{3} - \frac{1^4}{4}\right) - (0 - 0)\right] = 1$$

$$c\left(\frac{1}{3} - \frac{1}{4}\right) = 1$$

$$c\left(\frac{4-1}{12}\right) = 1$$

$$c\left(\frac{1}{12}\right) = 1$$

$$c = 12$$

Example 11:

A random variable x has the density function

$$f(x) = \begin{cases} \frac{1}{4}, & -2 < x < 2 \\ 0, & \text{else where} \end{cases}$$

obtain (i) P (–1 < x < 2) (ii) P (x >1)

Solution:

(i) $P(-1 < x < 2) = \int_{-1}^{2} f(x)\,dx$

$$\int_{-1}^{2} \frac{1}{4}\, dx = \frac{1}{4}\left[x\right]_{-1}^{+2}$$

$$= \frac{1}{4}\, [2 - (-1)]$$

$$= \frac{1}{4}\, [3]$$

$$= \frac{3}{4}$$

(ii) Here the upper limit of the p.d.f is $2 \therefore$ the probability for the given random variable.

$$P(\ x > 1) = \int_{1}^{2} \frac{1}{4}\, dx$$

$$= \frac{1}{4}\, [x]_{1}^{2}$$

$$= \frac{1}{4}\, [2 - 1]$$

$$= \frac{1}{4}\, [1]$$

$$= \frac{1}{4}$$

2.5 Mathematical Expectation:

Expectation is a very basic concept and is employed widely in decision theory, management science, system analysis, theory of games and many other fields. Some of these applications will be discussed in the chapter on Decision Theory.

The expected value or mathematical expectation of a random variable X is the weighted average of the values that X can assume with probabilities of its various values as weights.

Thus the expected value of a random variable is obtained by considering the various values that the variable can take multiplying these by their corresponding probabilities and summing these products. Expectation of X is denoted by E(X)

2.5.1 Expectation of a discrete random variable:

Let X be a discrete random variable which can assume any of the values of x_1, x_2, x_3.......x_n with respective probabilities p_1, p_2, p_3.....$p_{n.}$ Then the mathematical expectation of X is given by

$$E(x) = x_1p_1 + x_2p_2 + x_3p_3 + ...x_np_n$$

$$= \sum_{i=1}^{n} x_ip_i \text{ , where } \sum_{i=1}^{n} p_i = 1$$

Note:

Mathematical expectation of a random variable is also known as its arithmetic mean. We shall give some useful theorems on expectation without proof.

2.5.2 Theorems on Expectation:

1. For two random variable X and Y if E(X) and E(Y) exist, E(X + Y) = E(X) + E(Y) . This is known as addition theorem on expectation.
2. For two independent random variable X and Y, E(XY) = E(X).E(Y) provided all expectation exist. This is known as multiplication theorem on expectation.
3. The expectation of a constant is the constant it self.
   ie E(C) = C
4. E(cX) = cE(X)
5. E(aX+b) = aE(X) +b
6. Variance of constant is zero. ie Var(c) = 0
7. Var(X+c) = Var X
   Note: This theorem gives that variance is independent of change of origin.
8. Var (aX) = a^2 var(X)
   Note: This theorem gives that change of scale affects the variance.
9. Var (aX+b) = a^2Var(X)
10. Var (b-ax) = a^2 Var(x)

Definition:

Let f(x) be a function of random variable X. Then expectation of f(x) is given by E(f(x)) = Σ f(x) P(X=x) , where P(X=x) is the probability function of x.

Particular cases:

1. If we take f(x) = X^r, then $E(X^r) = \Sigma x^r p(x)$ is defined as the **r^{th} moment about origin** or r^{th} raw moment of the probability distribution. It is denoted by μ'_r

Thus $\mu'_r = E(X^r)$

$\mu'_1 = E(X)$

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

Hence mean = $\overline{X} = \mu'_1 = E(X)$

$$\text{Variance} = \frac{\Sigma X^2}{N} - \left[\frac{\Sigma X}{N}\right]^2$$

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

$$= \mu'_2 - (\mu'_1)^2$$

Variance is denoted by μ_2

2. If we take $f(x) = (X - \overline{X})^r$ then $E(X - \overline{X})^r = \Sigma(X - \overline{X})^r p(x)$ which is μ_r, the **r^{th} moment about mean or r^{th} central moment.**

In particular if r = 2, we get

$$\mu_2 = E(X - \overline{X})^2$$

$$= \Sigma(X - \overline{X})^2 p(X)$$

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

These two formulae give the variance of probability distribution in terms of expectations.

Example 12:

Find the expected value of x, where x represents the outcome when a die is thrown.

Solution:

Here each of the outcome (ie., number) 1, 2, 3, 4, 5 and 6 occurs with probability $\frac{1}{6}$. Thus the probability distribution of X will be

x	1	2	3	4	5	6
P(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Thus the expected value of X is

$$E(X) = \Sigma x_i p_i$$

$$= x_1p_1 + x_2p_2 + x_3p_3 + x_4p_4 + x_5p_5 + x_6p_6$$

$$E(X) = \left[1\times\frac{1}{6}\right] + \left[2\times\frac{1}{6}\right] + \left[3\times\frac{1}{6}\right] + \left[4\times\frac{1}{6}\right] + \left[5\times\frac{1}{6}\right] + \left[6\times\frac{1}{6}\right]$$

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

$$E(X) = 3.5$$

Remark:

In the games of chance, the expected value of the game is defined as the value of the game to the player.

The game is said to be favourable to the player if the expected value of the game is positive, and unfavourable, if value of the game is negative. The game is called a fair game if the expected value of the game is zero.

Example 13:

A player throws a fair die. If a prime number occurs he wins that number of rupees but if a non-prime number occurs he loses that number of rupees. Find the expected gain of the player and conclude.

Solution:

Here each of the six outcomes in throwing a die have been assigned certain amount of loss or gain. So to find the expected gain of the player, these assigned gains (loss is considered as negative gain) will be denoted as X.

These can be written as follows:

Outcome on a die	1	2	3	4	5	6
Associated gain to the outcome (x_i)	-1	2	3	- 4	5	- 6
P(x_i)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Note that 2,3 and 5 are prime numbers now the expected gain is

E (x)$\sum\sum_{i=1}^{6} x_i p_i$

$$= (-1)\left[\frac{1}{6}\right] + (2)\left[\frac{1}{6}\right] + (3)\left[\frac{1}{6}\right] + (-4)\left[\frac{1}{6}\right] + (5)\left[\frac{1}{6}\right] + (-6)\left[\frac{1}{6}\right]$$

$$= -\left[\frac{1}{6}\right]$$

Since the expected value of the game is negative, the game is unfavourable to the player.

Example 14:

An urn contains 7 white and 3 red balls. Two balls are drawn together at random from the urn. Find the expected number of white balls drawn.

Solution:

From the urn containing 7 white and 3 red balls, two balls can be drawn in $10C_2$ ways. Let X denote the number of white balls drawn, X can take the values 0, 1 and 2.

The probability distribution of X is obtained as follows:

P(0) = Probability that neither of two balls is white.

= Probability that both balls drawn are red.

$$= \frac{3C_2}{10C_2} = \frac{3\times2}{10\times9} = \frac{1}{15}$$

P(1) = Probability of getting 1 white and 1 red ball.

$$= \frac{7C_1\times3C_1}{10C_2} = \frac{7\times3\times2}{10\times9} = \frac{7}{15}$$

P(2) = Probability of getting two white balls

$$= \frac{7C_2}{10C_2} = \frac{7\times6}{10\times9} = \frac{7}{15}$$

Hence expected number of white balls drawn is

$$E(x) = \Sigma\, x_i\, p(x_i) = \left[0\times\frac{1}{15}\right] + \left[1\times\frac{7}{15}\right] + \left[2\times\frac{7}{15}\right]$$

$$= \frac{7}{5} = 1.4$$

Example 15:

A dealer in television sets estimates from his past experience the probabilities of his selling television sets in a day is given below. Find the expected number of sales in a day.

Number of TV Sold in a day	0	1	2	3	4	5	6
Probability	0.02	0.10	0.21	0.32	0.20	0.09	0.06

Solution:

We observe that the number of television sets sold in a day is a random variable which can assume the values 0, 1,2, 3,4,5,6 with the respective probabilities given in the table.

Now the expectation of x = $E(X) = \Sigma x_i p_i$

$= x_1p_1 + x_2p_2 + x_3p_3 + x_4p_4 + x_5p_5 + x_6p_6$

$= (0)(0.02) + (1)(0.010) + 2(0.21) + (3)(0.32) + 4(0.20) + (5)(0.09) + (6)(0.06)$

$E(X) = 3.09$

The expected number of sales per day is 3

Example 16:

Let x be a discrete random variable with the following probability distribution

X	-3	6	9
P(X= x)	1/6	1/2	1/3

Find the mean and variance

Solution:

$$E(x) = \Sigma x_i\, p_i$$

$$= (-3)\left[\frac{1}{6}\right] + (6)\left[\frac{1}{2}\right] + (9)\left[\frac{1}{3}\right]$$

$$= \left[\frac{11}{2}\right]$$

$E(x^2) \quad = \quad \Sigma\, x_i^2\, p_i$

$$= (-3)^2 \left[\frac{1}{6}\right] + (6)^2 \left[\frac{1}{2}\right] + (9)^2 \left[\frac{1}{3}\right] = \left[\frac{93}{2}\right]$$

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

$$= \left[\frac{93}{2}\right] - \left[\frac{11}{2}\right]^2$$

$$= \left[\frac{93}{2}\right] - \left[\frac{121}{4}\right]$$

$$= \frac{186-121}{4}$$

$$= \frac{65}{4}$$

2.5.3 Expectation of a continuous random variable:

Let X be a continuous random variable with probability density function f(x), then the mathematical expectation of x is defined as

E(x) = $\int_{-\infty}^{\infty} xf(x)dx$, provided the integral exists.

Remark:

If g(x) is function of a random variable and E[g(x)] exists,

then E[(g(x)] = $\int_{-\infty}^{\infty}$ g(x) f(x)dx

Example 17:

Let X be a continuous random variable with p.d.f given by f(x) = $4x^3$, 0 < x < 1. Find the expected value of X.

Solution:

We know that E(X) = $\int_{-\infty}^{\infty} xf(x)dx$

In this problem E(X) = $\int_{0}^{1} x(4x^3)dx$

$$= 4\int_0^1 x(x^3)dx$$

$$= 4\left[\frac{x^5}{5}\right]_0^1$$

$$= \frac{4}{5}\left[x^5\right]_0^1$$

$$= \frac{4}{5}\ [\ 1^5 - 0^5]$$

$$= \frac{4}{5}[1]$$

$$= \frac{4}{5}$$

Example 18:

Let x be a continuous random variable with pdf. given by $f(x) = 3x^2$, $0 < x < 1$ Find mean and variance

Solution:

$$E(x) \quad = \quad \int_{-\infty}^{\infty} xf(x)\,dx$$

$$E(x) \quad = \quad \int_0^1 x[3x^2]\,dx$$

$$= 3\int_0^1 x^3 dx$$

$$= 3\left[\frac{x^4}{4}\right]_0^1$$

$$= \frac{3}{4}\left[x^4\right]_0^1$$

$$= \frac{3}{4}\left[1^4 - 0\right]$$

$$= \frac{3}{4}$$

$$E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$$

$$= \int_0^1 x^2[3x^2]\,dx$$

$$= 3\int_0^1 x^4 dx$$

$$= 3\left[\frac{x^5}{5}\right]_0^1$$

$$= \frac{3}{5}\left[x^5\right]_0^1$$

$$= \frac{3}{5}\left[1^5 - 0\right]$$

$$= \frac{3}{5}$$

$$\text{Variance} = E(x^2) - [E(x)]^2$$

$$\text{Var }(x) = \frac{3}{5} - \left(\frac{3}{4}\right)^2$$

$$= \frac{3}{5} - \frac{9}{16}$$

$$= \frac{48-45}{80} = \frac{3}{80}$$

2.6 Moment generating function (M.G.F) (concepts only):

To find out the moments, the moment generating function is a good device. The moment generating function is a special form of mathematical expectation and is very useful in deriving the moments of a probability distribution.

Definition:

If X is a random variable, then the expected value of e^{tx} is known as the moment generating functions, provided the expected value exists for every value of t in an interval, - h < t < h , where h is some positive real value.

The moment generating function is denoted as $M_x(t)$
For discrete random variable

$$\begin{aligned} M_x(t) &= E(e^{tx}) \\ &= \Sigma\, e^{tx}\, p(x) \\ &= \Sigma \left(1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + ...\right) p_x(x) \end{aligned}$$

$$M_x(t) = \left(1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + ...\right) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r'$$

In the above expression, the r^{th} raw moment is the coefficient of $\frac{t^r}{r!}$ in the above expanded sum. To find out the moments differentiate the moment generating function with respect to t once, twice, thrice……and put t = 0 in the first, second, third, …. derivatives to obtain the first, second, third,……. moments.

From the resulting expression, we get the raw moments about the origin. The central moments are obtained by using the relationship between raw moments and central moments.

2.7 Characteristic function:

The moment generating function does not exist for every distribution. Hence another function, which always exists for all the distributions is known as characteristic function.

It is the expected value of e^{itx}, where $i = \sqrt{-1}$ and t has a real value and the characteristic function of a random variable X is denoted by $\phi_x(t)$

For a discrete variable X having the probability function p(x), the characteristic function is $\phi_x(t) = \Sigma\, e^{itx}\, p(x)$

For a continuous variable X having density function f(x), such that *a* < x < b , the characteristic function $\phi_x(t) = \int_a^b e^{itx} f(x)dx$

Exercise - 2

I. Choose the best answer:

1. $\sum_{i=1}^{n} p(x_i)$ is equal to

   (a) 0 (b) 1 (c) –1 (d) ∞

2. If F(x) is distribution function, then F(-∞) is

   (a) –1 (b) 0 (c) 1 (d) -∞

3. From the given random variable table, the value of *a* is

X=x	0	1	2
p_i	*a*	2*a*	*a*

   (a) 1 (b) $\frac{1}{2}$ (c) 4 (d) $\frac{1}{4}$

4. E(2x+3) is

   (a) E(2x) (b) 2E(x) +3 (c) E(3) (d) 2x+3

5. Var(x+8) is

   (a) var (8) (b) var(x) (c) 8 var(x) (d) 0

6. Var(5x+2) is

   (a) 25 var (x) (b) 5 var (x) (c) 2 var (x) (d) 25

7. Variance of the random variable X is

   (a) $E(x^2) - [E(x)]^2$ (b) $[E(x)]^2 - E(x^2)$

   (c) $E(x^2)$ (d) $[E(x)]^2$

8. Variance of the random variable x is $\frac{1}{16}$; its standard deviation is

   (a) $\frac{1}{256}$ (b) $\frac{1}{32}$

   (c) $\frac{1}{64}$ (d) $\frac{1}{4}$

9. A random variable X has E(x) = 2 and $E(x^2)$ = 8 its variance is

   (a) 4 (b) 6

   (c) 8 (d) 2

10. If f(x) is the p.d.f of the continuous random variable x, then $E(x^2)$ is

(a) $\int_{-\infty}^{\infty} f(x)\,dx$ (b) $\int_{-\infty}^{\infty} xf(x)\,dx$

(c) $\int_{-\infty}^{\infty} x^2 f(x)\,dx$ (d) $\int_{-\infty}^{\infty} f(x^2)\,dx$

II. Fill in the blanks:

11. If f(x) is a distribution function, then F(+∞) is equal to ________
12. If F(x) is a cumulative distribution function of a continuous random variable x with p.d.f f(x) then F′(x) = __________
13. f(x) is the probability density function of a continuous random variable X. Then $\int_{-\infty}^{\infty} f(x)\,dx$ is equal to ________
14. Mathematical expectation of a random variable X is also known as _____________
15. Variance of a constant is _____________
16. Var (12x) is _____________
17. Var (4x+7) is _________
18. If x is a discrete random variable with the probabilities p_i , then the expected value of x^2 is ________
19. If f(x) is the p.d.f of the continuous random variable X, then the expectation of X is given by __________
20. The moment generating function for the discrete random variable is given by ____________

III. Answer the following:

21. Define random variable.
22. Define discrete random variable
23. Define continuous random variable
24. What is probability mass function?
25. What is discrete probability distribution?
26. Define probability density function.

27. Write the properties of distribution function.
28. Define mathematical expectation for discrete random variable.
29. Define the expectation of a continuous random variable.
30. State the moment generating function.
31. State the characteristic function for a discrete random variable.
32. State the characteristic function for the continuous random variable.
33. Write short note on moment generating function.
34. Write a short note on characteristic function.
35. Find the probability distribution of X when 3 coins are tossed, where x is defined as getting head.
36. Two dice are thrown simultaneously and getting three is termed as success. Obtain the probability distribution of the number of threes.
37. Three cards are drawn at random successively, with replacement, from a well shuffled pack of 52 cards. Getting a card of diamond is termed as success. Obtain the probability distribution of the number of success.
38. A random variable X has the following probability distribution

Value of x	0	1	2	3	4
P(X=x)	$3a$	$4a$	$6a$	$7a$	$8a$

(a) determine the value of a (b) Find $p(1 < x < 4)$

(c) $P(1 \leq x \leq 4)$ (d) Find $P(x > 2)$

(e) Find the distribution function of x

39. A random variable X has the following probability function.

Values of X, x	0	1	2	3	4	5	6	7
P(x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

(i) Find k (ii) Find $p(0 < x < 5)$

(iii) Find $p(x \leq 6)$

40. Verify whether the following are probability density function

(i) $f(x) = 6x^5$, $0 < x < 1$

(ii) $f(x) = \frac{2x}{9}$, $0 < x < 3$

41. A continuous random variable x follows the probability law. $f(x) = Ax^3$, $0 < x < 1$ determine A

42. A random variable X has the density function $f(x) = 3x^2$, $0 < x < 1$ Find the probability between 0.2 and 0.5

43. A random variable X has the following probability distribution

X=x	5	2	1
P(x)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Find the expected value of x

44. A random variable X has the following distribution

x	-1	0	1	2
P(x)	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$

Find E(x) , $E(x^2)$ and Var (x)

45. A random variable X has $E(x) = \frac{1}{2}$ and $E(x^2) = \frac{1}{2}$ find its variance and standard deviation.

46. In a continuous distribution, whose probability density function is given by $f(x) = \frac{3}{4}x(2-x)$, $0 < x < 2$. Find the expected value of x.

47. The probability density function of a continuous random variable X is given by $f(x) = \frac{x}{2}$ for $0 < x < 2$. Find its mean and variance

Answers

I.

1. (b) 2. (b) 3. (d) 4. (b) 5. (b)

6. (a) 7. (a) 8. (d) 9. (a) 10. (c)

II.

11. 1 12. f(x) 13. 1 14. Mean

15. zero 16. 144 var(x) 17. 16 var(x)

18. $\Sigma x_i^2 \, p_i$ 19. $\int_{-\infty}^{\infty} x \, f(x) \, dx$ 20. $\sum_{r=0}^{\infty} \frac{t^r}{r!} \mu!_r$

III.

35.

X = x	0	1	2	3
$P(x_i)$	1/8	3/8	3/8	1/8

36.

X=x	0	1	2
P(x=x)	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$

37.

X = x	0	1	2	3
$P(x_i)$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

38. (i) *a* = 1/28 (ii) 13/28 (iii) 25/28 (iv) 15/28

(v)

x	0	1	2	3	4
F(x)	$\frac{3}{28}$	$\frac{7}{28}$	$\frac{13}{28}$	$\frac{20}{28}$	$\frac{28}{28}=1$

39. (i) k = 1/10 (ii) 4/5 (iii) 83/100 40. (i) p.d.f (ii) p.d.f

41. A = 4 42. P(0.2 < x , 0.5) = 0.117 43. 2.5

44. E(x) = 1/2 , var (x) = 19/12 45. 1/4 , 1/2 46. E(x) = 1

47. E(x) = 4/3 , var (x) = 2/9

3. SOME IMPORTANT THEORETICAL DISTRIBUTIONS

3.1 BINOMIAL DISTRIBUTION

3.1.0 Introduction:

In this chapter we will discuss the theoretical discrete distributions in which variables are distributed according to some definite probability law, which can be expressed mathematically. The Binomial distribution is a discrete distribution expressing the probability of a set of dichotomous alternative i.e., success or failure. This distribution has been used to describe a wide variety of process in business and social sciences as well as other areas.

3.1.1 Bernoulli Distribution:

A random variable X which takes two values 0 and 1 with probabilities q and p i.e., P(x=1) = p and P(x=0) = q, q = 1–p, is called a Bernoulli variate and is said to be a Bernoulli Distribution, where p and q takes the probabilities for success and failure respectively. It is discovered by Swiss Mathematician James Bernoulli (1654-1705).

Examples of Bernoulli's Trails are:

1) Toss of a coin (head or tail)
2) Throw of a die (even or odd number)
3) Performance of a student in an examination (pass or fail)

3.1.2 Binomial Distribution:

A random variable X is said to follow binomial distribution, if its probability mass function is given by

$$P(X = x) = P(x) = \begin{cases} nC_x \; p^x \, q^{n-x} & ; \quad x = 0, 1,2, \ldots,n \\ 0 & ; \quad \text{otherwise} \end{cases}$$

Here, the two independent constants n and p are known as the 'parameters' of the distribution. The distribution is completely determined if n and p are known. x refers the number of successes.

If we consider N sets of n independent trials, then the number of times we get x success is N(nC_x p^x q^{n-x}). It follows that the terms in the expansion of N $(q + p)^n$ gives the frequencies of the occurrences of 0,1,2,...,x,...,n success in the N sets of independent trials.

3.1.3 Condition for Binomial Distribution:

We get the Binomial distribution under the following experimental conditions.

1) The number of trials ‘ n’ is finite.
2) The trials are independent of each other.
3) The probability of success ‘ p’ is constant for each trial.
4) Each trial must result in a success or a failure.

The problems relating to tossing of coins or throwing of dice or drawing cards from a pack of cards with replacement lead to binomial probability distribution.

3.1.4 Characteristics of Binomial Distribution:

1. Binomial distribution is a discrete distribution in which the random variable X (the number of success) assumes the values 0,1, 2, …n, where n is finite.
2. Mean = np, variance = npq and

   standard deviation $\sigma = \sqrt{npq}$,

   Coefficient of skewness $= \dfrac{q-p}{\sqrt{npq}}$,

   Coefficient of kurtosis $= \dfrac{1 - 6pq}{npq}$, clearly each of the probabilities is non-negative and sum of all probabilities is 1 (p < 1 , q < 1 and p + q =1, q = 1– p).
3. The mode of the binomial distribution is that value of the variable which occurs with the largest probability. It may have either one or two modes.
4. If two independent random variables X and Y follow binomial distribution with parameter (n_1, p) and (n_2, p) respectively, then their sum (X+Y) also follows Binomial distribution with parameter ($n_1 + n_2$, p)

5. If n independent trials are repeated N times, N sets of n trials are obtained and the expected frequency of x success is N($nC_x\ p^x\ q^{n-x}$). The expected frequencies of 0,1,2…n success are the successive terms of the binomial distribution of $N(q + p)^n$.

Example 1:

Comment on the following: " The mean of a binomial distribution is 5 and its variance is 9"

Solution:

The parameters of the binomial distribution are n and p

We have mean ⇒ np = 5

Variance ⇒ npq = 9

$$\therefore q = \frac{npq}{np} = \frac{9}{5}$$

$$q = \frac{9}{5} > 1$$

Which is not admissible since q cannot exceed unity. Hence the given statement is wrong.

Example 2:

Eight coins are tossed simultaneously. Find the probability of getting atleast six heads.

Solution:

Here number of trials, n = 8, p denotes the probability of getting a head.

$$\therefore p = \frac{1}{2} \text{ and } q = \frac{1}{2}$$

If the random variable X denotes the number of heads, then the probability of a success in n trials is given by

$$P(X = x) = nc_x\, p^x\, q^{n-x}, \quad x = 0, 1, 2, ..., n$$

$$= 8C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} = 8C_x \left(\frac{1}{2}\right)^8$$

$$= \frac{1}{2^8}\, 8C_x$$

Probability of getting atleast six heads is given by

$$P(x \geq 6) = P(x = 6) + P(x = 7) + P(x = 8)$$

$$= \frac{1}{2^8}\ 8C_6 + \frac{1}{2^8}\ 8C_7 + \frac{1}{2^8}\ 8C_8$$

$$= \frac{1}{2^8}\ [\ 8C_6 + 8C_7 + 8C_8]$$

$$= \frac{1}{2^8}\ [\ 28 + 8 + 1] \quad = \frac{37}{256}$$

Example 3:

Ten coins are tossed simultaneously. Find the probability of getting (i) atleast seven heads (ii) exactly seven heads (iii) atmost seven heads

Solution:

$$p = \text{Probability of getting a head} \quad = \frac{1}{2}$$

$$q = \text{Probability of not getting a head} = \frac{1}{2}$$

The probability of getting x heads throwing 10 coins simultaneously is given by

$P(X = x) = nC_x\ p^x\ q^{n-x}.\quad ,\quad x = 0, 1, 2, ..., n$

$$= 10C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = \frac{1}{2^{10}}\ 10C_x$$

i) Probability of getting atleast seven heads

$$P(x \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$$

$$= \frac{1}{2^{10}}\ [\ 10C_7 + 10C_8 + 10C_9 + 10C_{10}]$$

$$= \frac{1}{1024}\ [\ 120 + 45 + 10 + 1] = \frac{176}{1024}$$

ii) Probability of getting exactly 7 heads

$$P(\ x = 7) = \frac{1}{2^{10}}\ 10C_7 = \frac{1}{2^{10}}\ (120)$$

$$= \frac{120}{1024}$$

iii) Probability of getting atmost 7 heads

$$\begin{aligned} P(x \leq 7) &= 1 - P(x > 7) \\ &= 1 - \{ P(x = 8) + P (x = 9) + P(x = 10)\} \\ &= 1 - \frac{1}{2^{10}} \{10C_8 + 10C_9 + 10C_{10}\} \\ &= 1 - \frac{1}{2^{10}} [45 +10+1] \\ &= 1 - \frac{56}{1024} \\ &= \frac{968}{1024} \end{aligned}$$

Example 4:

20 wrist watches in a box of 100 are defective. If 10 watches are selected at random, find the probability that (i) 10 are defective (ii) 10 are good (iii) at least one watch is defective (iv) at most 3 are defective.

Solution:

20 out of 100 wrist watches are defective

Probability of defective wrist watch , $p = \frac{20}{100} = \frac{1}{5}$

$$\therefore q = 1- p = \frac{4}{5}$$

Since 10 watches are selected at random, n =10

$P(X = x) = nC_x \, p^x q^{n-x}$, x = 0 , 1, 2, ...,10

$$= 10C_x \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{10-x}$$

i) Probability of selecting 10 defective watches

$$\begin{aligned} P(x =10) &= 10C_{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^0 \\ &= 1. \frac{1}{5^{10}} .1 = \frac{1}{5^{10}} \end{aligned}$$

ii) Probability of selecting 10 good watches (i.e. no defective)

$$P(x = 0) = 10C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10}$$

$$= 1.1. \left(\frac{4}{5}\right)^{10} = \left(\frac{4}{5}\right)^{10}$$

iii) Probability of selecting at least one defective watch

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

$$= 1 - P(x = 0)$$

$$= 1 - 10C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10}$$

$$= 1 - \left(\frac{4}{5}\right)^{10}$$

iv) Probability of selecting at most 3 defective watches

$$P (x \leq 3) = P (x = 0) + P(x =1) + P(x = 2) + P(x = 3)$$

$$= 10C_0\left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10} +10C_1\left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 + 10C_2\left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8$$

$$+10C_3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7$$

$$= 1.1. \left(\frac{4}{5}\right)^{10} + 10\left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 + \frac{10.9}{1.2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8$$

$$+ \frac{10.9.8}{1.2.3} \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7$$

$$= 1. (0.107) + 10 (0.026) + 45 (0.0062) + 120 (0.0016)$$

$$= 0.859 \text{ (approx)}$$

Example 5:

With the usual notation find p for binomial random variable X if n = 6 and 9P(X = 4) = P(X = 2)

Solution:

The probability mass function of binomial random variable X is given by

$P(X = x) = nC_x\, p^x\, q^{n-x}.$, $x = 0 , 1, 2, ...,n$

Here n = 6 $\therefore$ $P(X = x) = 6C_x\, p^x q^{6-x}$

$P(x = 4) = 6C_4\, p^4 q^2$

$P(x = 2) = 6C_2\, p^2 q^4$

Given that,

$9.\ P(x = 4) = P(x = 2)$

$9.\ 6C_4\, p^4q^2 = 6C_2\, p^2q^4$

$\Rightarrow 9 \times 15p^2 = 15q^2$

$9p^2 = q^2$

Taking positive square root on both sides we get,

$$3p = q$$
$$= 1- p$$
$$4p = 1$$
$$\therefore\ p = \frac{1}{4} = 0.25$$

3.1.5 Fitting of Binomial Distribution:

When a binomial distribution is to be fitted to an observed data, the following procedure is adopted.

1. Find Mean $= \bar{x} = \frac{\Sigma fx}{\Sigma f} = np$

   $\Rightarrow p = \frac{\bar{x}}{n}$ where n is number of trials
2. Determine the value, $q = 1- p$.
3. The probability function is $P(x) = {}_nC_x\, p^x q^{n-x}$ put x = 0, we set $P(0) = q^n$ and $f(0) = N \times P(0)$
4. The other expected frequencies are obtained by using the recurrence formula is given by

   $f(x+1) = \frac{n-x}{x+1}\ \frac{p}{q}\ f(x)$

Example 6:

A set of three similar coins are tossed 100 times with the following results

Number of heads :	0	1	2	3
Frequency :	36	40	22	2

Solution:

X	f	fx
0	36	0
1	40	40
2	22	44
3	2	6
	Σf =100	Σfx = 90

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{90}{100} = 0.9$$

$$p = \frac{\bar{x}}{n}$$

$$= \frac{0.9}{3} = 0.3$$

$$q = 1 - 0.3$$

$$= 0.7$$

The probability function is $P(x) = nC_x\, p^x\, q^{n-x}$

Here n = 3, p = 0.3 q = 0.7

$$\therefore P(x) = 3C_x\, (0.3)^x\, (0.7)^{3-x}$$

$$P(0) = 3C_0\, (0.3)^0\, (0.7)^3$$

$$= (0.7)^3 \quad = 0.343$$

$\therefore$ f(0) = N × P(0) = 0.343 × 100 = 34.3

The other frequencies are obtained by using the recurrence formula $f(x+1) = \frac{n-x}{x+1}\left(\frac{p}{q}\right) f(x)$. By putting x = 0, 1, 2 the expected frequencies are calculated as follows.

$$f(1) = \frac{3-0}{0+1}\left(\frac{p}{q}\right) \times 34.3$$

$$= 3 \times (0.43) \times 34.3 = 44.247$$

$$f(2) = \frac{3-1}{1+1}\left(\frac{p}{q}\right) f(1)$$

$$= \frac{2}{2}\,(0.43) \times 44.247$$

$$= 19.03$$

$$f(3) = \frac{3-2}{2+1}\left(\frac{p}{q}\right) f(2)$$

$$= \frac{1}{3}(0.43) \times 19.03$$

$$= 2.727$$

The observed and theoretical (expected) frequencies are tabulated below:

					Total
Observed frequencies	36	40	22	2	100
Expected frequencies	34	44	19	3	100

Example 7:

4 coins are tossed and number of heads noted. The experiment is repeated 200 times and the following distribution is obtained .

x: Number of heads	0	1	2	3	4
f: frequencies	62	85	40	11	2

Solution:

X	0	1	2	3	4	Total
f	62	85	40	11	2	200
fx	0	85	80	33	8	206

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{206}{200} = 1.03$$

$$p = \frac{\bar{x}}{n} = \frac{1.03}{4} = 0.2575$$

$$\therefore\ q = 1 - 0.2575 = 0.7425$$

Here n = 4 , p = 0.2575 ; q = 0.7425

The probability function of binomial distribution is

$$P(x) = nC_x\ p^x\, q^{n-x}$$

The binomial probability function is

$$P(x) = 4C_x \ (0.2575)^x \ (0.7425)^{4-x}$$

$$P(0) = (0.7425)^4$$

$$= 0.3039$$

$$\therefore \ f(0) = NP(0)$$

$$= 200 \times 0.3039$$

$$= 60.78$$

The other frequencies are calculated using the recurrence formula $f(x+1) = \frac{n-x}{x+1}\left(\frac{p}{q}\right) f(x)$. By putting x = 0,1, 2, 3 then the expected frequencies are calculated as follows:

Put x = 0, we get

$$f(1) = \frac{4-0}{0+1}(0.3468)\ (60.78)$$

$$= \ 84.3140$$

$$f(2) = \frac{4-1}{1+1}\ (0.3468)\ (84.3140)$$

$$= \ 43.8601$$

$$f(3) = \frac{4-2}{2+1}\ (0.3468)\ (43.8601)$$

$$= 10.1394$$

$$f(4) = \frac{4-3}{3+1}\ (0.3468)\ (10.1394)$$

$$= \ 0.8791$$

The theoretical and expected frequencies are tabulated below:

						Total
Observed frequencies	62	85	40	11	2	200
Expected frequencies	61	84	44	10	1	200

3.2 POISSON DISTRIBUTION:

3.2.0 Introduction:

Poisson distribution was discovered by a French Mathematician-cum-Physicist Simeon Denis Poisson in 1837. Poisson distribution is also a discrete distribution. He derived it as a limiting case of Binomial distribution. For n-trials the binomial distribution is $(q + p)^n$; the probability of x successes is given by $P(X=x) = nC_x\, p^x\, q^{n-x}$. If the number of trials n is very large and the probability of success ' p' is very small so that the product np = m is non – negative and finite.

The probability of *x* success is given by

$$P(\ X = x\ ) = \begin{cases} \dfrac{e^{-m}\, m^x}{x!} & \text{for } x = 0,1,2, \ldots \\ 0 & ;\ \text{otherwise} \end{cases}$$

Here m is known as parameter of the distribution so that m >0

Since number of trials is very large and the probability of success p is very small, it is clear that the event is a rare event. Therefore Poisson distribution relates to rare events.

Note:

1) e is given by $e = 1 + \dfrac{1}{1!} + \dfrac{1}{2!} + \dfrac{1}{3!} + \ldots = 2.71828$

2) $P(X=0) = \dfrac{e^{-m}\, m^0}{0!}$, $0! = 1$ and $1! = 1$

3) $P(X=1) = \dfrac{e^{-m}\, m^1}{1!}$

Some examples of Poisson variates are :

1. The number of blinds born in a town in a particular year.
2. Number of mistakes committed in a typed page.
3. The number of students scoring very high marks in all subjects
4. The number of plane accidents in a particular week.
5. The number of defective screws in a box of 100, manufactured by a reputed company.
6. Number of suicides reported in a particular day.

3.2.1 Conditions:

Poisson distribution is the limiting case of binomial distribution under the following conditions:

1. The number of trials n is indefinitely large i.e., n à ∞
2. The probability of success ‘ p’ for each trial is very small; i.e., p à 0
3. np = m (say) is finite , m > 0

3.2.2 Characteristics of Poisson Distribution:

The following are the characteristics of Poisson distribution

1. Discrete distribution: Poisson distribution is a discrete distribution like Binomial distribution, where the random variable assume as a countably infinite number of values 0,1,2 ….
2. The values of p and q: It is applied in situation where the probability of success p of an event is very small and that of failure q is very high almost equal to 1 and n is very large.
3. The parameter: The parameter of the Poisson distribution is m. If the value of m is known, all the probabilities of the Poisson distribution can be ascertained.
4. Values of Constant: Mean = m = variance; so that standard deviation = $\sqrt{m}$

   Poisson distribution may have either one or two modes.
5. Additive Property: If X and Y are two independent Poisson distribution with parameter m_1 and m_2 respectively. Then (X+Y) also follows the Poisson distribution with parameter $(m_1 + m_2)$
6. As an approximation to binomial distribution: Poisson distribution can be taken as a limiting form of Binomial distribution when n is large and p is very small in such a way that product np = m remains constant.
7. Assumptions: The Poisson distribution is based on the following assumptions.
   i) The occurrence or non- occurrence of an event does not influence the occurrence or non-occurrence of any other event.

ii) The probability of success for a short time interval or a small region of space is proportional to the length of the time interval or space as the case may be.

iii) The probability of the happening of more than one event is a very small interval is negligible.

Example 8:

Suppose on an average 1 house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district, what is the probability that exactly 5 houses will have a fire during the year? [given that $e^{-2} = 0.13534$]

Mean, $\bar{x} = np$, $n = 2000$ and $p = \frac{1}{1000}$

$$= 2000 \times \frac{1}{1000}$$

$$m = 2$$

The Poisson distribution is

$$P(X=x) = \frac{e^{-m} m^x}{x!}$$

$$\therefore P(X=5) = \frac{e^{-2} 2^5}{5!}$$

$$= \frac{(0.13534)\times 32}{120}$$

$$= 0.036$$

(Note: The values of e^{-m} are given in Appendix)

Example 9:

In a Poisson distribution 3P(X=2) = P(X=4) Find the parameter ‘ m’ .

Solution:

Poisson distribution is given by $P(X=x) = \frac{e^{-m} m^x}{x!}$

Given that 3P(x=2) = P(x= 4)

$$3.\ \frac{e^{-m} m^2}{2!} = \frac{e^{-m} m^4}{4!}$$

$$m^2 = \frac{3 \times 4!}{2!}$$

$$\therefore \quad m = \pm 6$$

Since mean is always positive $\therefore$ m = 6

Example 10:

If 2% of electric bulbs manufactured by a certain company are defective. Find the probability that in a sample of 200 bulbs i) less than 2 bulbs ii) more than 3 bulbs are defective.[e^{-4} = 0.0183]

Solution:

The probability of a defective bulb = p = $\frac{2}{100}$ = 0.02

Given that n = 200 since p is small and n is large

We use the Poisson distribution

mean, m = np = 200 × 0.02 = 4

Now, Poisson Probability function, P(X = x) = $\frac{e^{-m} m^x}{x!}$

i) Probability of less than 2 bulbs are defective

$$= P(X<2)$$
$$= P(x = 0) + P(x = 1)$$
$$= \frac{e^{-4} 4^0}{0!} + \frac{e^{-4} 4^1}{1!}$$
$$= e^{-4} + e^{-4} (4)$$
$$= e^{-4} (1 + 4) = 0.0183 \times 5$$
$$= 0.0915$$

ii) Probability of getting more than 3 defective bulbs

$$P(x > 3) = 1 - P(x \leq 3)$$
$$= 1 - \{P(x = 0) + P(x = 1) + P(x=2) + P(x=3)\}$$
$$= 1 - e^{-4} \{1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\}$$
$$= 1 - \{0.0183 \times (1 + 4 + 8 + 10.67)\}$$
$$= 0.567$$

3.2.3 Fitting of Poisson Distribution:

The process of fitting of Poisson distribution for the probabilities of x = 0, 1,2,... success are given below :

i) First we have to calculate the mean = $\bar{x} = \dfrac{\sum fx}{\sum f} = m$

ii) The value of e^{-m} is obtained from the table (see Appendix)

iii) By using the formula $P(X=x) = \dfrac{e^{-m}.m^{x}}{x!}$

Substituting x = 0, $P(0) = e^{-m}$

Then $f(0) = N \times P(0)$

The other expected frequencies will be obtained by using the recurrence formula

$$f(x+1) = \frac{m}{x+1} f(x) \text{ ; } x = 0,1,2, \ldots$$

Example 11:

The following mistakes per page were observed in a book.

Number of mistakes (per page)	0	1	2	3	4
Number of pages	211	90	19	5	0

Fit a Poisson distribution to the above data.

Solution:

x_i	f_i	f_ix_i
0	211	0
1	90	90
2	19	38
3	5	15
4	0	0
	N = 325	Σfx = 143

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{N}$$

$$= \frac{143}{325} = 0\ .44 = m$$

Then $e^{-m} \Rightarrow e^{-0.44} = 0.6440$

Probability mass function of Poisson distribution is

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

Put x = 0, $P(0) = e^{-0.44}\,\dfrac{44^{0}}{0!}$

$$= e^{-0.44}$$

$$= 0.6440$$

$$\therefore f(0) = N\,P(0)$$

$$= 325 \times 0.6440$$

$$= 209.43$$

The other expected frequencies will be obtained by using the recurrence formula

$f(x+1) = \dfrac{m}{x+1} f(x)$. By putting x = 0,1,2,3 we get the expected frequencies and are calculated as follows.

$$f(1) = 0.44 \times 209.43 = 92.15$$

$$f(2) = \frac{0.44}{2} \times 92.15 = 20.27$$

$$f(3) = \frac{0.44}{3} \times 20.27 = 2.97$$

$$f(4) = \frac{0.44}{4} \times 2.97 = 0.33$$

						Total
Observed frequencies	211	90	19	5	0	325
Expected frequencies	210	92	20	3	0	325

Example 12:

Find mean and variance to the following data which gives the frequency of the number of deaths due to horse kick in 10 corps per army per annum over twenty years.

X	0	1	2	3	4	Total
F	109	65	22	3	1	200

Solution:

Let us calculate the mean and variance of the given data

x_i	f_i	$f_i x_i$	$f_i x_i^2$
0	109	0	0
1	65	65	65
2	22	44	88
3	3	9	27
4	1	4	16
Total	N = 200	$\Sigma fx = 122$	$\Sigma fx^2 = 196$

$$\text{Mean} = \bar{x} = \frac{\sum_i f_i x}{N}$$

$$= \frac{122}{200}$$

$$= 0.61$$

$$\text{Variance} = \sigma^2 = \frac{\sum_i f_i^2 x}{N} - \left(\bar{x}\right)^2$$

$$= \frac{196}{200} - (0.61)^2$$

$$= 0.61$$

Hence, mean = variance = 0.61

Example 13:

100 car radios are inspected as they come off the production line and number of defects per set is recorded below

No. of defects	0	1	2	3	4
No. of sets	79	18	2	1	0

Fit a Poisson distribution and find expected frequencies

Solution:

x	f	f*x*
0	79	0
1	18	18
2	2	4
3	1	3
4	0	0
	N = 100	Σ*fx* = 25

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{N}$$

$$= \frac{25}{100}$$

$$\therefore m = 0.25$$

Then $e^{-m} = e^{-0.25} = 0.7788 = 0.779$

Poisson probability function is given by

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

$$P(0) = \frac{e^{-0.25}(0.25)^0}{0!} = (0.779)$$

$\therefore$ f(0) = N.P(0) = 100 × (0.779) = 77.9

Other frequencies are calculated using the recurrence formula

$$f(x+1) = \frac{m}{x+1} f(x).$$

By putting x = 0,1,2,3, we get the expected frequencies and are calculated as follows.

$$f(1) = f(0+1) = \frac{m}{0+1} f(0)$$

$$f(1) = \frac{0.25}{1}(77.9)$$

$$= 19.46$$

$$f(2) = \frac{0.25}{2}(19.46)$$

$$= 2.43$$

$$f(3) = \frac{0.25}{3}\ (2.43)$$
$$= 0.203$$
$$f(4) = \frac{0.25}{4}\ (0.203)$$
$$= 0.013$$

Observed frequencies	79	18	2	1	0	100
Expected frequencies	78	20	2	0	0	100

Example 14:

Assuming that one in 80 births in a case of twins, calculate the probability of 2 or more sets of twins on a day when 30 births occurs. Compare the results obtained by using (i) the binomial and (ii) Poisson distribution.

Solution:

(i) Using Binomial distribution

Probability of twins birth $= p = \frac{1}{80} = 0.0125$

$$\therefore\ q = 1 - p = 1 - 0.0125$$
$$= 0.9875$$
$$n = 30$$

Binomial distribution is given by

$$P(x) = nC_x\ p^x\ q^{n-x}$$
$$P(x \geq 2) = 1 - P(x < 2)$$
$$= 1 - \{P(x = 0) + P(x = 1)\}$$
$$= 1 - \{30C_0(0.0125)^0 (0.9875)^{30} + 30C_1 (0.0125)^1 (0.9875)^{29}\}$$
$$= 1 - \{1.1(0.9875)^{30} + 3\ (0.125)\ (0.9875)^{29}\}$$
$$= 1 - \{\ 0.6839 + 0.2597\}$$
$$= 1 - 0.9436$$
$$P(\ x \geq 2) = 0.0564$$

(ii) By using Poisson distribution:

The probability mass function of Poisson distribution is given by

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

Mean = m = np

= 30 (0.0125) = 0.375

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

$= 1 - \{ P(x = 0) + P(x = 1)\}$

$$= 1 - \{ \frac{e^{-0.375}(0.375)^{0}}{0!} + \frac{e^{-0.375}(0.375)^{1}}{1!} \}$$

$= 1 - e^{-0.375} (1 + 0.375)$

$= 1 - (0.6873)(1.375) = 1 - 0.945 = 0.055$

3.3 NORMAL DISTRIBUTION:

3.3.0 Introduction:

In the preceding sections we have discussed the discrete distributions, the Binomial and Poisson distribution.

In this section we deal with the most important continuous distribution, known as normal probability distribution or simply normal distribution. It is important for the reason that it plays a vital role in the theoretical and applied statistics.

The normal distribution was first discovered by DeMoivre (English Mathematician) in 1733 as limiting case of binomial distribution. Later it was applied in natural and social science by Laplace (French Mathematician) in 1777. The normal distribution is also known as Gaussian distribution in honour of Karl Friedrich Gauss(1809).

3.3.1 Definition:

A continuous random variable X is said to follow normal distribution with mean μ and standard deviation σ, if its probability density function

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

Note:

The mean μ and standard deviation σ are called the parameters of Normal distribution. The normal distribution is expressed by X ~ N(μ, σ^2)

3.3.2 Condition of Normal Distribution:

i) Normal distribution is a limiting form of the binomial distribution under the following conditions.

a) n, the number of trials is indefinitely large ie., n à ∞ and

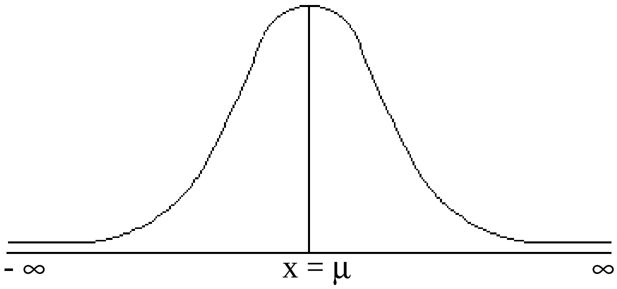
b) Neither p nor q is very small.

ii) Normal distribution can also be obtained as a limiting form of Poisson distribution with parameter m à ∞

iii) Constants of normal distribution are mean = μ, variation =σ^2, Standard deviation = σ.

3.3.3 Normal probability curve:

The curve representing the normal distribution is called the normal probability curve. The curve is symmetrical about the mean (μ), bell-shaped and the two tails on the right and left sides of the mean extends to the infinity. The shape of the curve is shown in the following figure.



3.3.4 Properties of normal distribution:

1. The normal curve is bell shaped and is symmetric at $x = \mu$.
2. Mean, median, and mode of the distribution are coincide i.e., Mean = Median = Mode = μ
3. It has only one mode at $x = \mu$ (i.e., unimodal)
4. Since the curve is symmetrical, Skewness = $\beta_1 = 0$ and Kurtosis = $\beta_2 = 3$.
5. The points of inflection are at $x = \mu \pm \sigma$
6. The maximum ordinate occurs at $x = \mu$ and its value is $= \dfrac{1}{\sigma\sqrt{2\pi}}$
7. The x axis is an asymptote to the curve (i.e. the curve continues to approach but never touches the x axis)
8. The first and third quartiles are equidistant from median.
9. The mean deviation about mean is 0.8 σ
10. Quartile deviation = 0.6745 σ
11. If X and Y are independent normal variates with mean μ_1 and μ_2, and variance σ_1^2 and σ_2^2 respectively then their sum (X + Y) is also a normal variate with mean $(\mu_1 + \mu_2)$ and variance $(\sigma_1^2 + \sigma_2^2)$
12. Area Property $P(\mu - \sigma < x < \mu + \sigma) = 0.6826$
$P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.9544$
$P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.9973$

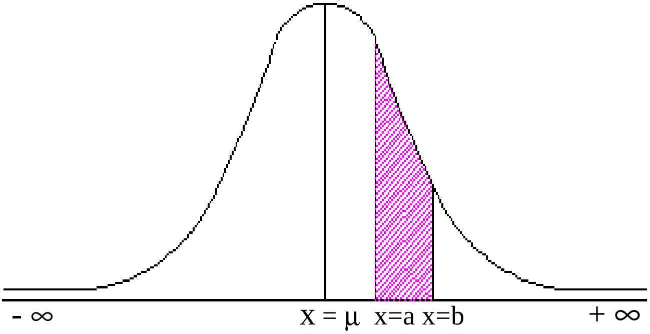
3.3.5 Standard Normal distribution:

Let X be random variable which follows normal distribution with mean μ and variance σ^2 .The standard normal variate is defined as $Z = \dfrac{X - \mu}{\sigma}$ which follows standard normal distribution with mean 0 and standard deviation 1 i.e., Z ~ N(0,1). The standard normal distribution is given by $\phi(z) = \dfrac{1}{\sqrt{2\pi}}\ e^{\frac{-1}{2}z^2}$; $-\infty < z < \infty$

The advantage of the above function is that it doesn’ t contain any parameter. This enable us to compute the area under the normal probability curve.

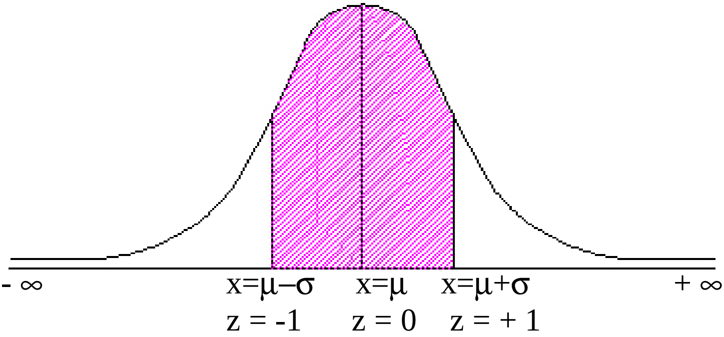
3.3.6 Area properties of Normal curve:

The total area under the normal probability curve is 1. The curve is also called standard probability curve. The area under the curve between the ordinates at x = a and x = b where a < b, represents the probabilities that x lies between x = a and x = b i.e., $P(a \leq x \leq b)$



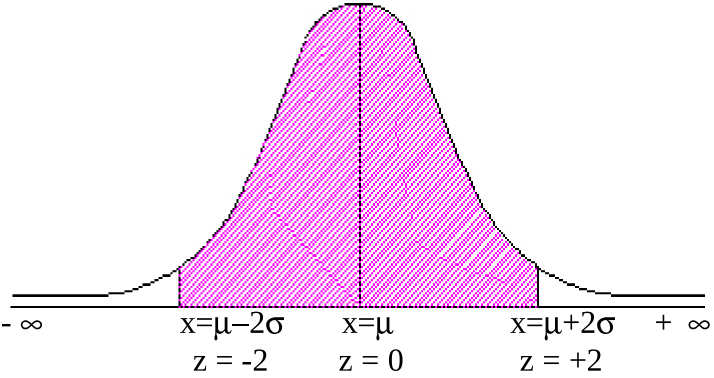
To find any probability value of x, we first standardize it by using $Z = \frac{X-\mu}{\sigma}$, and use the area probability normal table. (given in the Appendix).

For Example: The probability that the normal random variable x to lie in the interval $(\mu-\sigma , \mu+\sigma)$ is given by

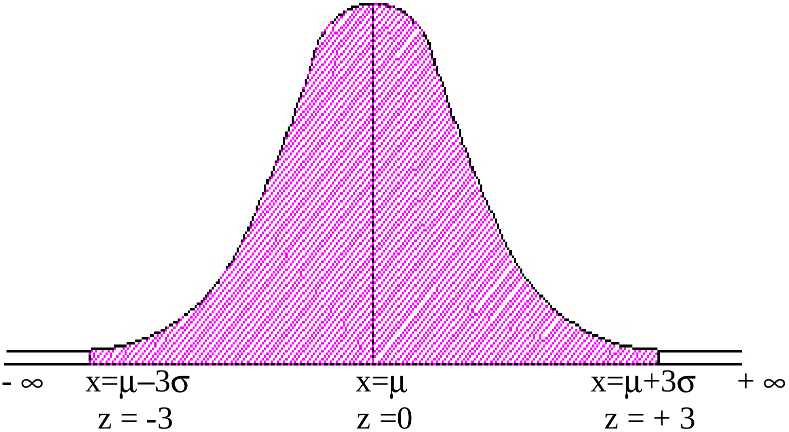


$$\begin{aligned} P(\mu - \sigma < x < \mu + \sigma) &= P(-1 \leq z \leq 1) \\ &= 2P(0 < z < 1) \\ &= 2\,(0.3413) \quad \text{(from the area table)} \\ &= 0.6826 \end{aligned}$$

$$\begin{aligned} P(\mu - 2\sigma < x < \mu + 2\sigma) &= P(-2 < z < 2) \\ &= 2P(0 < z < 2) \\ &= 2\,(0.4772) = 0.9544 \end{aligned}$$



$$\begin{aligned} P(\mu - 3\sigma < x < \mu + 3\sigma) &= P(-3 < z < 3) \\ &= 2P(0 < z < 3) \\ &= 2\,(0.49865) = 0.9973 \end{aligned}$$



The probability that a normal variate x lies outside the range $\mu \pm 3\sigma$ is given by

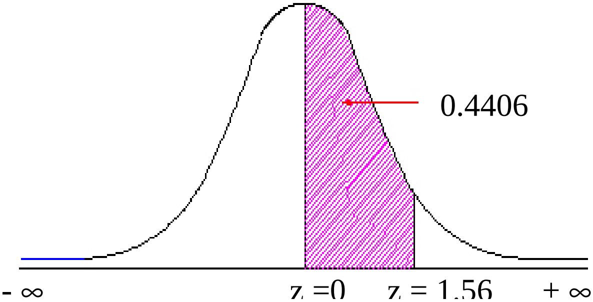
$$P(|x - \mu| > 3\sigma) = P(|z| > 3)$$
$$= 1 - P(-3 \le z \le 3)$$
$$= 1 - 0.9773 = 0.0027$$

Thus we expect that the values in a normal probability curve will lie between the range $\mu \pm 3\sigma$, though theoretically it range from – ∞ to ∞.

Example 15:

Find the probability that the standard normal variate lies between 0 and 1.56

Solution:

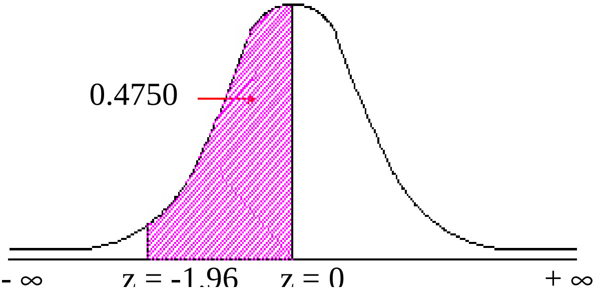


$$P(0<z<1.56) = \text{Area between } z = 0 \text{ and } z = 1.56$$
$$= 0.4406 \quad \text{(from table)}$$

Example 16:

Find the area of the standard normal variate from –1.96 to 0.

Solution:



Area between z = 0 & z =1.96 is same as the area z = –1.96 to z = 0

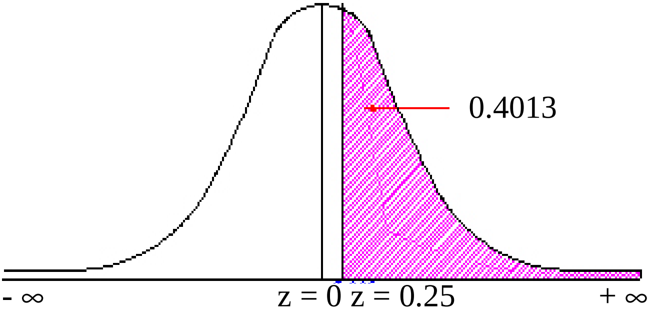
$P(-1.96 < z < 0) = P(0 < z < 1.96)$ (by symmetry)

$= 0.4750$ (from the table)

Example 17:

Find the area to the right of z = 0.25

Solution:



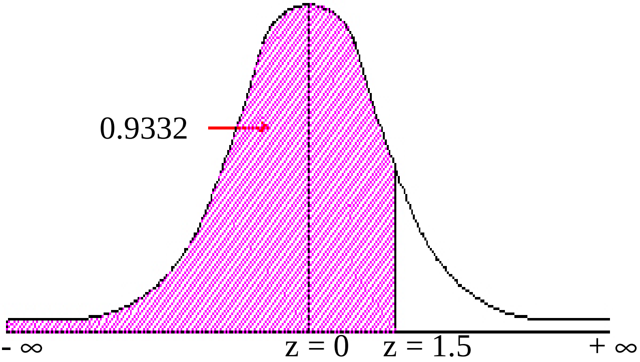
$P(z > 0.25) = P(0<z<\infty) - P(0<z<0.25)$

$= 0.5000 - 0.0987$ (from the table) $= 0.4013$

Example 18:

Find the area to the left of z = 1.5

Solution:



$P(z < 1.5) = P(-\infty < z < 0) + P(0 < z < 1.5)$

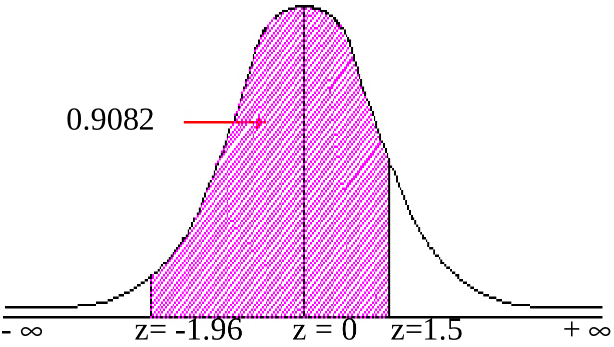
$= 0.5 + 0.4332$ (from the table)

$= 0.9332$

Example 19:

Find the area of the standard normal variate between –1.96 and 1.5

Solution:

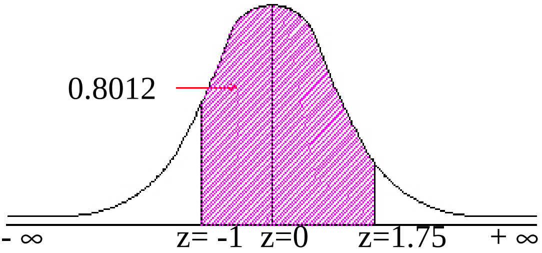


$$P(-1.96 < z < 1.5) = P(-1.96 < z < 0) + P(0 < z < 1.5)$$
$$= P(0 < z < 1.96) + P(0 < z < 1.5)$$
$$= 0.4750 + 0.4332 \quad \text{(from the table)}$$
$$= 0.9082$$

Example 20:

Given a normal distribution with $\mu = 50$ and $\sigma = 8$, find the probability that x assumes a value between 42 and 64

Solution:



Given that $\mu = 50$ and $\sigma = 8$

The standard normal variate $z = \dfrac{x-\mu}{\sigma}$

If X = 42 , $Z_1 = \dfrac{42-50}{8} = \dfrac{-8}{8} = -1$

If X = 64, $Z_2 = \dfrac{64-50}{8} = \dfrac{14}{8} = 1.75$

$\therefore$ P(42 < x < 64) = P(–1 < z <1.75)

= P(–1< z < 0) + P(0 < z <1.95)

= P(0<z<1) + P (0 < z <1.75) (by symmetry)

= 0.3413 +0 .4599 (from the table)

= 0 .8012

Example 21:

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

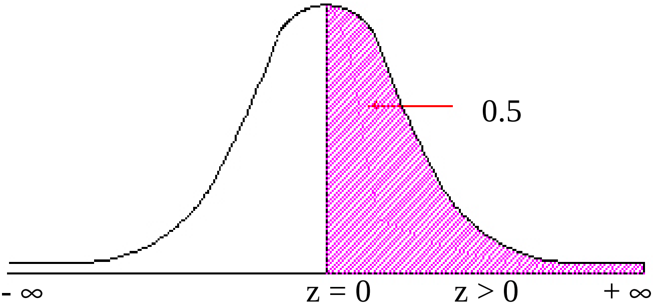
i) More than 60 marks (ii) Less than 56 marks

(iii) Between 45 and 65 marks

Solution:

Given that mean = μ = 60 and standard deviation = σ = 5

i) The standard normal varaiate $Z = \dfrac{x-\mu}{\sigma}$



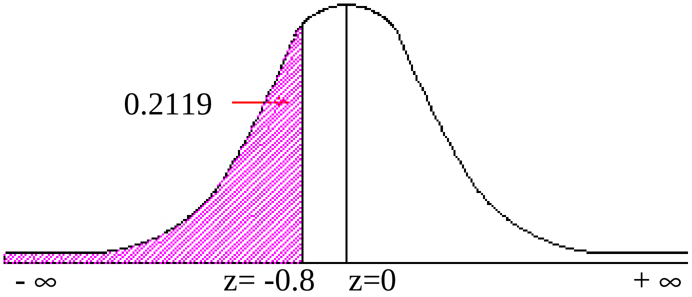
If X = 60, $Z = \dfrac{x-\mu}{\sigma} = \dfrac{60-60}{5} = 0$

$\therefore$P(x > 60) = P(z > 0)

= P(0 < z < ∞) = 0.5000

Hence the percentage of students scored more than 60 marks is 0.5000(100) = 50 %

ii) If X = 56, Z = $\frac{56-60}{5} = \frac{-4}{5} = -0.8$



P(x < 56) = P(z < –0.8)

= P(- ∞ < z < 0) – P(–0.8 < z < 0) (by symmetry)

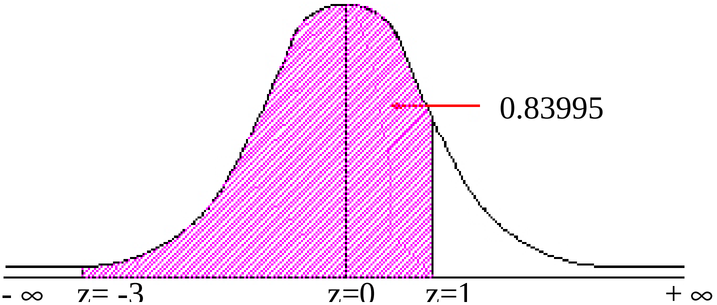
= P(0 < 2 < ∞) – P(0 < z < 0.8)

= 0.5 – 0.2881 (from the table)

= 0.2119

Hence the percentage of students score less than 56 marks is 0.2119(100) = 21.19 %

iii) If X = 45, then z = $\frac{45-60}{5} = \frac{-15}{5} = -3$



X = 65 then z = $\frac{65-60}{5} = \frac{5}{5} = 1$

P(45 < x < 65) = P(–3 < z < 1)

= P(–3 < z < 0) + P (0 < z < 1)

$= P(0 < z < 3) + P(0 < z < 1)$ (by symmetry)

$= 0.4986 + 0.3413$ (from the table)

$= 0.8399$

Hence the percentage of students scored between 45 and 65 marks is 0.8399(100) = 83.99 %

Example 22:

X is normal distribution with mean 2 and standard deviation 3. Find the value of the variable x such that the probability of the interval from mean to that value is 0.4115

Solution:

Given $\mu = 2$, $\sigma = 3$

Suppose z_1 is required standard value,

Thus $P(0 < z < z_1) = 0.4115$

From the table the value corresponding to the area 0.4115 is 1.35 that is $z_1 = 1.35$

Here $z_1 = \dfrac{x - \mu}{\sigma}$

$$1.35 = \frac{x - 2}{3}$$

$$x = 3(1.35) + 2$$

$$= 4.05 + 2 = 6.05$$

Example 23:

In a normal distribution 31 % of the items are under 45 and 8 % are over 64. Find the mean and variance of the distribution.

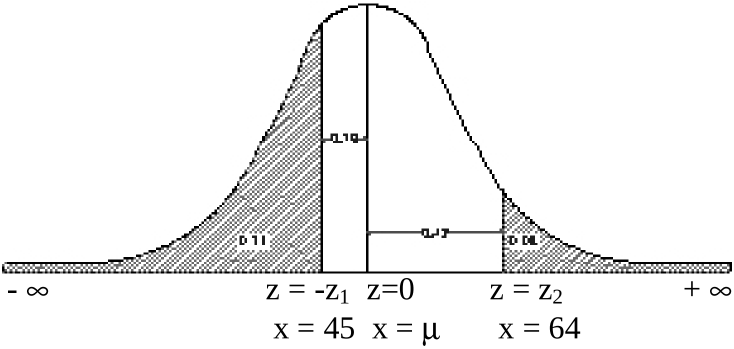
Solution:

Let x denotes the items are given and it follows the normal distribution with mean μ and standard deviation σ

The points x = 45 and x = 64 are located as shown in the figure.

i) Since 31 % of items are under x = 45, position of x into the left of the ordinate $x = \mu$

ii) Since 8 % of items are above x = 64 , position of this x is to the right of ordinate $x = \mu$



When x = 45, z = $\dfrac{x-\mu}{\sigma} = \dfrac{45-\mu}{\sigma} = -\, z_1$ (say)

Since x is left of x = μ , z_1 is taken as negative

When x = 64, z = $\dfrac{64-\mu}{\sigma}$ = z_2 (say)

From the diagram P(x < 45) = 0.31

$P(z < -\, z_1) = 0.31$

$P(-\, z_1 < z < 0) = P(-\infty < z < 0) - p(-\infty < z < z_1)$

s $= 0.5 - 0.31 = 0.19$

$P(0 < z < z_1) = 0.19$ (by symmetry)

$z_1 = 0.50$ (from the table)

Also from the diagram p(x > 64) = 0.08

$P(0 < z < z_2) = P(0 < z < \infty) - P(z_2 < z < \infty)$

$= 0.5 - 0.08 = 0.42$

$z_2 = 1.40$ (from the table)

Substituting the values of z_1 and z_2 we get

$\dfrac{45-\mu}{\sigma} = -\ 0.50$ and $\dfrac{64-\mu}{\sigma} = 1.40$

Solving $\mu - 0.50\,\sigma = 45$ ----- (1)

$\mu + 1.40\,\sigma = 64$ ----- (2)

(2) – (1) $\Rightarrow 1.90\,\sigma = 19 \Rightarrow \sigma = 10$

Substituting $\sigma = 10$ in (1) $\mu = 45 + 0.50\,(10)$

$= 45 + 5.0 = 50.0$

Hence mean = 50 and variance = $\sigma^2 = 100$

Exercise – 3

I. Choose the best answer:

1. Binomial distribution applies to

   (a) rare events (b) repeated alternatives

   (c) three events (d) impossible events

2. For Bernoulli distribution with probability p of a success and q of a failure, the relation between mean and variance that hold is

   (a) mean < variance (b) mean > variance

   (c) mean = variance (d) mean $\leq$ variance

3. The variance of a binomial distribution is

   (a) npq (b) np (c) $\sqrt{npq}$ (d) 0

4. The mean of the binomial distribution 15Cx $\left(\frac{2}{3}\right)^{x}\left(\frac{1}{3}\right)^{15-x}$ in which $p = \frac{2}{3}$ is

   (a) 5 (b) 10 (c) 15 (d) 3

5. The mean and variance of a binomial distribution are 8 and 4 respectively. Then P(x = 1) is equal to

   (a) $\frac{1}{2^{12}}$ (b) $\frac{1}{2^{4}}$ (c) $\frac{1}{2^{6}}$ (d) $\frac{1}{2^{8}}$

6. If for a binomial distribution , n = 4 and also P(x = 2) = 3P(x=3) then the value of p is

   (a) $\frac{9}{11}$ (b) 1 (c) $\frac{1}{3}$ (d) None of the above

7. The mean of a binomial distribution is 10 and the number of trials is 30 then probability of failure of an event is

   (a) 0.25 (b) 0.333 (c) 0.666 (d) 0.9

8. The variance of a binomial distribution is 2. Its standard deviation is

(a) 2 (b) 4 (c) 1/2 (d) $\sqrt{2}$

9. In a binomial distribution if the numbers of independent trials is n, then the probability of n success is

(a) $nC_xp^xq^{n-x}$ (b) 1 (c) p^n (d)q^n

10. The binomial distribution is completely determined if it is known

(a) p only (b) q only (c) p and q (d) p and n

11. The trials in a binomial distribution are

(a) mutually exclusive (b) non-mutually exclusive

(c) independent (d) non-independent

12. If two independent variables x and y follow binomial distribution with parameters,(n_1, p) and (n_2, p) respectively, their sum(x+y) follows binomial distribution with parameters

(a) $(n_1 + n_2, 2p)$ (b) (n, p)

(c) $(n_1 + n_2, p)$ (d) $(n_1 + n_2, p + q)$

13. For a Poisson distribution

(a) mean > variance (b) mean = variance

(c) mean < variance (d) mean $\leq$ variance

14. Poisson distribution correspondents to

(a) rare events (b) certain event

(c) impossible event (d) almost sure event

15. If the Poisson variables X and Y have parameters m_1 and m_2 then X+Y is a Poisson variable with parameter.

(a) m_1m_2 (b) m_1+m_2 (c) $m_1–m_2$ (d)m_1/m_2

16. Poisson distribution is a

(a) Continuous distribution

(b) discrete distribution

(c) either continuous or discrete

(d) neither continue nor discrete

17. Poisson distribution is a limiting case of Binomial distribution when

(a) n à ∞ ; p à 0 and np = $\sqrt{m}$

(b) n à 0 ; p à ∞ and p=1/m

(c) n à ∞ ; p à ∞ and np=m

(d) n à ∞ ; p à 0 ,np=m

18. If the expectation of a Poisson variable (mean) is 1 then P(x < 1) is

(a) e^{-1} (b) $1-2e^{-1}$

(c) $1- 5/2e^{-1}$ (d) none of these

19. The normal distribution is a limiting form of Binomial distribution if

(a) n à ∞ p à 0 (b) n à 0 , p à q (c) n à ∞ , p à n

(d) n à ∞ and neither p nor q is small.

20. In normal distribution, skewness is

(a) one (b) zero

(c) greater than one (d) less than one

21. Mode of the normal distribution is

(a) σ (b) $\frac{1}{\sqrt{2\pi}}$ (c) μ (d) 0

22. The standard normal distribution is represented by

(a) N(0,0) (b) N(1,1) (c) N(1,0) (d) N(0,1)

23. Total area under the normal probability curve is

(a) less than one (b) unity (c) greater than one (d) zero

24. The probability that a random variable x lies in the interval (μ - 2σ , μ + 2σ) is

(a) 0.9544 (b) 0.6826 (c) 0.9973 (d) 0.0027

25. The area P(- ∞ < z < 0) is equal to

(a) 1 (b) 0.1 (c) 0.5 (d) 0

26. The standard normal distribution has

(a) μ =1, σ = 0 (b) μ = 0, σ = 1 (c) μ = 0 ,σ = 0

(d) μ =1, σ = 1

27. The random variable x follows the normal distribution $f(x) = C.e^{-\frac{1}{2}\frac{(x-100)^2}{25}}$ then the value of C is

(a) $5\sqrt{2\pi}$ (b) $\frac{1}{5\sqrt{2\pi}}$ (c) $\frac{1}{\sqrt{2\pi}}$ (d) 5

28. Normal distribution has

(a) no mode (b) only one mode

(c) two modes (d) many mode

29. For the normal distribution

(a) mean = median =mode (b) mean < median < mode

(c) mean > median > mode (d) mean > median < mode

30. Probability density function of normal variable $P(X = x) = \frac{1}{5\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-30)^2}{25}}$; $-\alpha < x < \alpha$ then mean and variance are

(a) mean = 30 variance = 5 (b) mean = 0, variance = 25

(c) mean = 30 variance = 25 (d) mean = 30, variance = 10

31. The mean of a Normal distribution is 60, its mode will be

(a) 60 (b) 40 (c) 50 (d) 30

32. If x is a normal variable with μ =100 and σ^2 = 25 then P(90 < x < 120) is same as

(a) P(-1 < z < 1) (b) P(-2 < z < 4)

(c) P(4 < z < 4.1) (d) P(-2 < z < 3)

33. If x is N(6, 1.2) and $P(0 \leq z \leq 1) = 0.3413$ then $P(4.8 \leq x \leq 7.2)$ is

(a) 0.3413 (b) 0.6587 (c) 0.6826 (d) 0.3174

II. Fill in the blanks:

34. The probability of getting a head in successive throws of a coin is __________

35. If the mean of a binomial distribution is 4 and the variance is 2 then the parameter is ___________

36. $\left(\frac{2}{3}+\frac{1}{3}\right)^9$ refers the binomial distribution and its standard deviation is _________

37. In a binomial distribution if number of trials to be large and probability of success be zero, then the distribution becomes ________.

38. The mean and variance are _______ in Poisson distribution

39. The mean of Poisson distribution is 0.49 and its standard deviation is ________

40. In Poisson distribution, the recurrence formula to calculate expected frequencies is ______.

41. The formula $\frac{\sum fx^2}{N} - \left(\bar{x}\right)^2$ is used to find ________

42. In a normal distribution, mean takes the values from __________to ________

43. When $\mu = 0$ and $\sigma = 1$ the normal distribution is called ________

44. $P(-\infty < z < 0)$ covers the area ______

45. If $\mu = 1200$ and $\sigma = 400$ then the standard normal variate z for $x = 800$ is _________

46. At $x = \mu \pm \sigma$ are called as __________ in a normal distribution.

47. $P(-3 < z < 3)$ takes the value __________

48. X axis be the ________to the normal curve.

III. Answer the following

49. Comment the following

“ For a binomial distribution mean = 7 and variance = 16

50. Find the binomial distribution whose mean is 3 and variance 2

51. In a binomial distribution the mean and standard deviation are 12 and 2 respectively. Find n and p

52. A pair of dice is thrown 4 times. If getting a doublet is considered a success, find the probability of 2 success.

53. Explain a binomial distribution.

54. State the characteristics of a binomial distribution.
55. State the conditions for a binomial variate.
56. Explain the fitting of a binomial distribution.
57. For the binomial distribution $(0.68+0.32)^{10}$ find the probability of 2 success.
58. Find the mean of binomial distribution of the probability of occurrence of an event is 1/5 and the total number of trials is 100
59. If on an average 8 ships out of 10 arrive safely at a port, find the mean and standard deviation of the number of ships arriving safely out of total of 1600 ships.
60. The probability of the evening college student will be a graduate is 0.4. Determine the probability that out of 5 students (i) none (ii) one (iii) atleast one will be a graduate
61. Four coins are tossed simultaneously. What is the probability of getting i) 2 heads and 2 tails ii) atleast 2 heads iii) atleast one head.
62. 10% of the screws manufactured by an automatic machine are found to be defective. 20 screws are selected at random. Find the probability that i) exactly 2 are defective ii) atmost 3 are defective iii) atleast 2 are defective.
63. 5 dice are thrown together 96 times. The numbers of getting 4, 5 or 6 in the experiment is given below. Calculate the expected frequencies and compare the standard deviation of the expected frequencies and observed frequencies.

Getting 4 ,5 or 6 :	0	1	2	3	4	5
Frequency	: 1	10	24	35	18	8

64. Fit a binomial distribution for the following data and find the expected frequencies.

X :	0	1	2	3	4
f	18	35	30	13	4

65. Eight coins are tossed together 256 times. Number of heads observed at each toss is recorded and the results are given

below. Find the expected frequencies. What are the theoretical value of mean and standard deviation? Calculate also mean and standard deviation of the observed frequencies.

Number of heads:	0	1	2	3	4	5	6	7	8
Frequencies :	2	6	39	52	67	56	32	10	1

66. Explain Poisson distribution.
67. Give any two examples of Poisson distribution.
68. State the characteristics of Poisson distribution.
69. Explain the fitting of a Poisson distribution
70. A variable x follows a Poisson distribution with mean 6 calculate i) P(x = 0) ii) P(x = 2)
71. The variance of a Poisson Distribution is 0.5. Find P(x = 3). [$e^{-0.5}$ = 0.6065]
72. If a random variable X follows Poisson distribution such that P(x =1) = P(x = 2) find (a) the mean of the distribution and P(x = 0). [e^{-2} = 0.1353]
73. If 3% of bulbs manufactured by a company are defective then find the probability in a sample of 100 bulbs exactly five bulbs are defective.
74. It is known from the past experience that in a certain plant there are on the average 4 industrial accidents per month. Find the probability that in a given year there will be less than 3 accidents. Assume Poisson distribution.[e^{-4} = 0.0183]
75. A manufacturer of television sets known that of an average 5% of this product is defective. He sells television sets in consignment of 100 and guarantees that not more than 4 sets will be defective. What is the probability that a television set will fail to meet the guaranteed quality? [e^{-5} = 0.0067]
76. One fifth percent of the blades produced by a blade manufacturing factory turns out to be a defective. The blades are supplied in pockets of 10. Use Poisson distribution to calculate the approximate number of pockets containing i) no defective (ii) all defective (iii) two defective blades respectively in a consignment of 1,00,000 pockets.

77. A factory employing a huge number of workers find that over a period of time, average absentee rate is three workers per shift. Calculate the probability that in a given shift
i) exactly 2 workers (ii) more than 4 workers will be absent.

78. A manufacturer who produces medicine bottles finds that 0.1 % of the bottles are defective. They are packed in boxes containing 500 bottles. A drag manufactures buy 100 boxes from the producer of bottles. Using Poisson distribution find how many boxes will contain (i) no defective ii) exactly 2 (iii) atleast 2 defective.

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

Mistakes per page:	0	1	2	3	4	5
No of pages :	142	156	69	57	5	1

Fit a Poisson distribution.

80. Fit a Poisson distribution to the following data:

x:	0	1	2	3	4	5	6	7	8	Total
f:	229	325	257	119	50	17	2	1	0	1000

81. The following tables given that number of days in a 50, days period during which automatically accidents occurred in city. Fit a Poisson distribution to the data

No of accidents :	0	1	2	3	4
No of days :	21	18	7	3	1

82. Find the probability that standard normal variate lies between 0.78 and 2.75

83. Find the area under the normal curve between z = 0 and z = 1.75

84. Find the area under the normal curve between z = -1.5 and z = 2.6

85. Find the area to the left side of z = 1.96

86. Find the area under the normal curve which lies to the right of z = 2.70

87. A normal distribution has mean = 50 and standard deviation is 8. Find the probability that x assumes a value between 34 and 62

88. A normal distribution has mean = 20 and S.D = 10. Find area between x =15 and x = 40

89. Given a normal curve with mean 30 and standard deviation 5. Find the area under the curve between 26 and 40
90. The customer accounts of a certain departmental store have an average balance of Rs.1200 and a standard deviation of Rs.400. Assuming that the account balances are normally distributed. (i) what percentage of the accounts is over Rs.1500? (ii) What percentage of the accounts is between Rs.1000 and Rs.1500? iii) What percentage of the accounts is below Rs.1500?
91. The weekly remuneration paid to 100 lecturers coaching for professional entrance examinations are normally distributed with mean Rs.700 and standard deviation Rs.50. Estimate the number of lecturers whose remuneration will be i) between Rs.700 and Rs.720 ii) more than Rs.750 iii) less than Rs.630
92. x is normally distributed with mean 12 and standard deviation 4. Find the probability of the following i) $x \geq 20$ ii) $x \leq 20$ iii) $0 < x < 12$
93. A sample of 100 dry cells tested to find the length of life produced the following results μ =12 hrs, σ = 3 hrs. Assuming the data, to be normally distributed. What percentage of battery cells are expressed to have a life. i) more than 15 hrs ii) between 10 and 14 hrs as iii) less than 6 hrs?.
94. Find the mean and standard deviation of marks in an examination where 44 % of the candidates obtained marks below 55 and 6 % got above 80 marks.
95. In a normal distribution 7 % of the items are under 35 and 89 % of the items are under 63. Find its mean as standard deviation.

Note: For fitting a binomial distribution in the problem itself, if it is given that the coin is unbiased, male and female births are equally probable, then we consider p = q = ½. All other cases we have to find the value of p from the mean value of the given data.

Answers

I.

1. b 2. b 3. a 4. b 5. a

6. c 7. c 8. d 9. c 10. d

11.c 12.c 13. b 14. a 15. b

16. b 17. d 18. a 19. d 20. b

21. c 22. d 23. b 24. a 25. c

26. b 27. b 28. b 29. a 30. c

31. a 32. b 33. c 34. $\frac{1}{2}$ 35. $(8, \frac{1}{2})$

36. $\sqrt{2}$ 37. Poisson distribution 38. equal 39. 0.7

40. $f(x+1) = \frac{m}{x+1} f(x)$ 41. variance 42. $- \infty, + \infty$

43. Standard normal distribution 44. 0.5 45. –1

46. Point of inflections 47. 0.9973

48. Asymptote

49. This is not admissible . Since $q = \frac{16}{7} > 1$

50. $\left(\frac{2}{3}+\frac{1}{3}\right)^9$, $p = \frac{2}{3}$, $q = \frac{1}{3}$ and n = 9

51. n = 18, $p = \frac{2}{3}$. 52. $\frac{25}{216}$

57. $10\ C_2\ (0.32)^2+(0.68)^8$ 58. 20 59. 1280

60. i) 0.08 ii) 0.259 iii) 0.92 61. i) $\frac{3}{8}$ ii) $\frac{11}{16}$ iii) $\frac{15}{16}$

62. (i) $190 \times \frac{9^{18}}{10^{20}}$ (ii) $\frac{1}{10^{20}}\ [9^{20} + 20 \times 9^{19} + 190 \times 9^{18} + 1140 \times 9^{17}]$

(iii) $1- \frac{1}{10^{20}}\ [9^{20} + 20 \times 9^{19} + 190 \times 9^{18}]$

63. Observed S.D. = 1.13 and expected S.D.= 1.12

65. Observed mean = 4.0625 and S.D. = 1.462

70. i).0.00279 ii) 0.938 71. 0.0126

72. a) Mean = 2 b) P(x=0) = 0.1353

73. P(x = 5) = 0.1008 74. 0.2379 75. 0.9598

76. i) 98,020 ii)1960 iii) 20 77. i) 0.2241 ii) 0.1846

78. i) 61 ii) 76 iii) 9 79. P(x) = $\dfrac{e^{-1}\ 1^{x}}{x!}$

80. P(x) = $\dfrac{e^{-1.5}\,(1.5)^{x}}{x!}$ 81. P(x) = $\dfrac{e^{-0.9}\,(0.9)^{x}}{x!}$

82. 0.2147 83. 0.4599

84. 0.9285 85. 0.9750

86. 0.0035 87. 0.9104

88. 0.6687 89. 0.7653

90. i) 22.66 % ii) 46.49 % iii) 77.34 %

91. i) 16 ii)16 iii) 8

92. i) 0.0228 ii) 0.9772 iii) 0.4987

93. i) 15.87 % ii) 49.72 % iii) 2.28 %

94. Mean = 57.21 and SD = 14.71

95. Mean = 50.27 and SD = 10.35

4. TEST OF SIGNIFICANCE (Basic Concepts)

4.0 Introduction:

It is not easy to collect all the information about population and also it is not possible to study the characteristics of the entire population (finite or infinite) due to time factor, cost factor and other constraints. Thus we need sample. Sample is a finite subset of statistical individuals in a population and the number of individuals in a sample is called the sample size.

Sampling is quite often used in our day-to-day practical life. For example in a shop we assess the quality of rice, wheat or any other commodity by taking a handful of it from the bag and then to decide to purchase it or not.

4.1 Parameter and Statistic:

The statistical constants of the population such as mean, (μ), variance (σ^2), correlation coefficient (ρ) and proportion (P) are called ' Parameters' .

Statistical constants computed from the samples corresponding to the parameters namely mean ($\overline{x}$), variance (S^2), sample correlation coefficient (r) and proportion (p) etc, are called statistic.

Parameters are functions of the population values while statistic are functions of the sample observations. In general, population parameters are unknown and sample statistics are used as their estimates.

4.2 Sampling Distribution:

The distribution of all possible values which can be assumed by some statistic measured from samples of same size ' n' randomly drawn from the same population of size N, is called as sampling distribution of the statistic (DANIEL and FERREL).

Consider a population with N values .Let us take a random sample of size n from this population, then there are

$NC_n = \dfrac{N!}{n!(N-n)!}$ = k (say), possible samples. From each of these k samples if we compute a statistic (e.g mean, variance, correlation coefficient, skewness etc) and then we form a frequency distribution for these k values of a statistic. Such a distribution is called sampling distribution of that statistic.

For example, we can compute some statistic t = t(x_1, x_2,….x_n) for each of these k samples. Then t_1, t_2 …., t_k determine the sampling distribution of the statistic t. In other words statistic *t* may be regarded as a random variable which can take the values t_1, t_2 …., t_k and we can compute various statistical constants like mean, variance, skewness, kurtosis etc., for this sampling distribution.

The mean of the sampling distribution t is

$$\bar{t} = \frac{1}{K}[t_1 + t_2 + \ldots\ldots + t_k] = \frac{1}{K}\sum_{i=1}^{k} t_i$$

and var (t) = $\dfrac{1}{K}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots\ldots\ldots + (t_k - \bar{t})^2\right]$

$$= \frac{1}{K}\Sigma(t_i - \bar{t})^2$$

4.3 Standard Error:

The standard deviation of the sampling distribution of a statistic is known as its standard error. It is abbreviated as S.E. For example, the standard deviation of the sampling distribution of the mean $\bar{x}$ known as the standard error of the mean,

Where v($\bar{x}$) = v $\left(\dfrac{x_1 + x_2 + \ldots\ldots\ldots\ldots x_n}{n}\right)$

$$= \frac{v(x_1)}{n^2} + \frac{v(x_2)}{n^2} + \ldots\ldots. + \frac{v(x_n)}{n^2}$$

$$= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \ldots\ldots + \frac{\sigma^2}{n^2} = \frac{n\sigma^2}{n^2}$$

∴ The S.E. of the mean is $\dfrac{\sigma}{\sqrt{n}}$

The standard errors of the some of the well known statistic for large samples are given below, where *n* is the sample size, σ^2 is the population variance and *P* is the population proportion and Q = 1–P. n_1 and n_2 represent the sizes of two independent random samples respectively.

Sl.No	Statistic	Standard Error
1.	Sample mean $\overline{x}$	$\frac{\sigma}{\sqrt{n}}$
2.	Observed sample proportion p	$\sqrt{\frac{PQ}{n}}$
3.	Difference between of two samples means ($\overline{x}_1 - \overline{x}_2$)	$\sqrt{\frac{\sigma_{1_1}{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}$
4.	Difference of two sample proportions $p_1 - p_2$	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

Uses of standard error

i) Standard error plays a very important role in the large sample theory and forms the basis of the testing of hypothesis.

ii) The magnitude of the S.E gives an index of the precision of the estimate of the parameter.

iii) The reciprocal of the S.E is taken as the measure of reliability or precision of the sample.

iv) S.E enables us to determine the probable limits within which the population parameter may be expected to lie.

Remark:

S.E of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost, labour and time etc.,

4.4 Null Hypothesis and Alternative Hypothesis

Hypothesis testing begins with an assumption called a Hypothesis, that we make about a population parameter. A hypothesis is a supposition made as a basis for reasoning. The conventional approach to hypothesis testing is not to construct a single hypothesis about the population parameter but rather to set up two different hypothesis. So that of one hypothesis is accepted, the other is rejected and vice versa.

Null Hypothesis:

A hypothesis of no difference is called null hypothesis and is usually denoted by H_0 " Null hypothesis is the hypothesis" which is tested for possible rejection under the assumption that it is true " by Prof. R.A. Fisher. It is very useful tool in test of significance. For example: If we want to find out whether the special classes (for Hr. Sec. Students) after school hours has benefited the students or not. We shall set up a null hypothesis that "H_0: special classes after school hours has not benefited the students".

Alternative Hypothesis:

Any hypothesis, which is complementary to the null hypothesis, is called an alternative hypothesis, usually denoted by H_1, For example, if we want to test the null hypothesis that the population has a specified mean μ_0 (say),

i.e., : Step 1: null hypothesis H_0: $\mu = \mu_0$

then 2. Alternative hypothesis may be

i) $H_1 : \mu \neq \mu_0$ (ie $\mu > \mu_0$ or $\mu < \mu_0$)

ii) $H_1 : \mu > \mu_0$

iii) $H_1 : \mu < \mu_0$

the alternative hypothesis in (i) is known as a two – tailed alternative and the alternative in (ii) is known as right-tailed (iii) is known as left –tailed alternative respectively. The settings of alternative hypothesis is very important since it enables us to decide whether we have to use a single – tailed (right or left) or two tailed test.

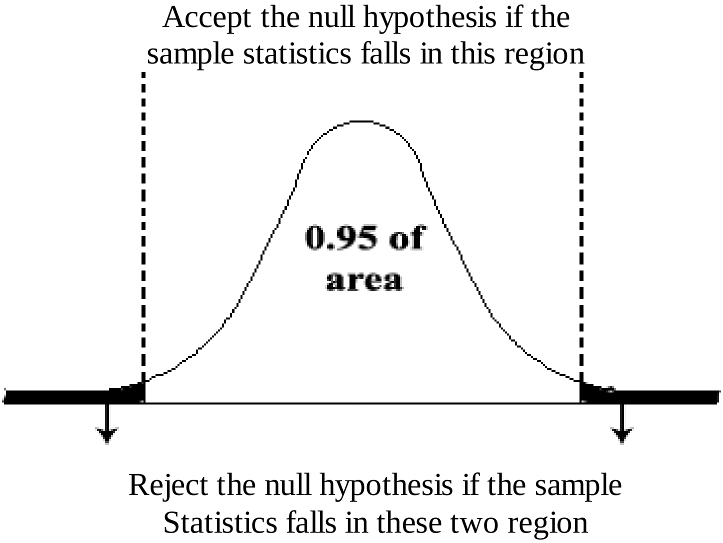
4.5 Level of significance and Critical value:

Level of significance:

In testing a given hypothesis, the maximum probability with which we would be willing to take risk is called level of significance of the test. This probability often denoted by "α" is generally specified before samples are drawn.

The level of significance usually employed in testing of significance are 0.05(or 5 %) and 0.01 (or 1 %). If for example a 0.05 or 5 % level of significance is chosen in deriving a test of hypothesis, then there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted. (i.e.,) we are about 95 % confident that we made the right decision. In such a case we say that the hypothesis has been rejected at 5 % level of significance which means that we could be wrong with probability 0.05.

The following diagram illustrates the region in which we could accept or reject the null hypothesis when it is being tested at 5 % level of significance and a two-tailed test is employed.


Accept the null hypothesis if the
sample statistics falls in this region
0.95 of
area
Reject the null hypothesis if the sample
Statistics falls in these two region

Note: Critical Region: A region in the sample space S which amounts to rejection of H_0 is termed as critical region or region of rejection.

Critical Value:

The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depends upon i) the level of significance used and ii) the alternative hypothesis, whether it is two-tailed or single-tailed

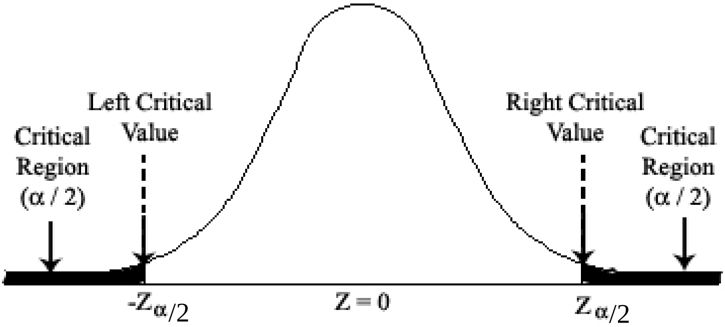
For large samples the standard normal variate corresponding to the statistic t, Z = $\left|\frac{t - E(t)}{S.E.(t)}\right|$ ~ N (0,1)

asymptotically as n à ∞

The value of z under the null hypothesis is known as test statistic. The critical value of the test statistic at the level of significance α for a two - tailed test is given by $Z_{\alpha/2}$ and for a one tailed test by Z_α. where Z_α is determined by equation $P(|Z| > Z_\alpha) = \alpha$

Zα is the value so that the total area of the critical region on both tails is α . ∴ $P(Z > Z_\alpha) = \frac{\alpha}{2}$. Area of each tail is $\frac{\alpha}{2}$.

Zα is the value such that area to the right of Zα and to the left of – Zα is $\frac{\alpha}{2}$ as shown in the following diagram.



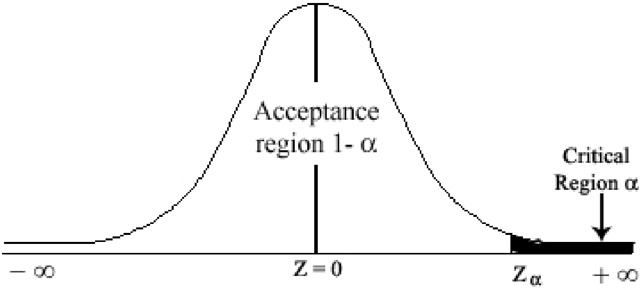
4.6 One tailed and Two Tailed tests:

In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

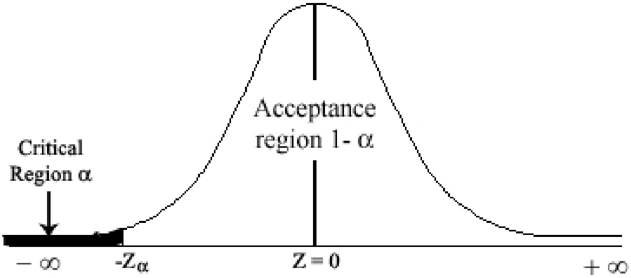
One tailed test: A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called a one tailed test.

For example, for testing the mean of a population H_0: $\mu = \mu_0$, against the alternative hypothesis H_1: $\mu > \mu_0$ (right – tailed) or $H_1 : \mu < \mu_0$ (left –tailed)is a single tailed test. In the right – tailed test H_1: $\mu > \mu_0$ the critical region lies entirely in right tail of the sampling distribution of $\bar{x}$, while for the left tailed test H_1: $\mu < \mu_0$ the critical region is entirely in the left of the distribution of $\bar{x}$.

Right tailed test:



Left tailed test:



Two tailed test:

A test of statistical hypothesis where the alternative hypothesis is two tailed such as, $H_0 : \mu = \mu_0$ against the alternative hypothesis H_1: $\mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is known as two tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of test of statistic.

For example, suppose that there are two population brands of washing machines, are manufactured by standard process(with mean warranty period μ_1) and the other manufactured by some new technique (with mean warranty period μ_2): If we want to test if the washing machines differ significantly then our null hypothesis is $H_0 : \mu_1 = \mu_2$ and alternative will be H_1: $\mu_1 \neq \mu_2$ thus giving us a two tailed test. However if we want to test whether the average warranty period produced by some new technique is more than those produced by standard process, then we have $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$ thus giving us a left-tailed test.

Similarly, for testing if the product of new process is inferior to that of standard process then we have, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1>\mu_2$ thus giving us a right-tailed test. Thus the decision about applying a two – tailed test or a single –tailed (right or left) test will depend on the problem under study.

Critical values ($Z\alpha$) of Z

Level of significance α	0.05 or 5%		0.01 or 1%	
	Left	Right	Left	Right
Critical values of Z_α for one tailed Tests	–1.645	1.645	–2.33	2.33
Critical values of $Z_{\alpha/2}$ for two tailed tests	–1.96	1.96	–2.58	2.58

4.7 Type I and Type II Errors:

When a statistical hypothesis is tested there are four possibilities.

1. The hypothesis is true but our test rejects it (Type I error)
2. The hypothesis is false but our test accepts it (Type II error)
3. The hypothesis is true and our test accepts it (correct decision)
4. The hypothesis is false and our test rejects it (correct decision)

Obviously, the first two possibilities lead to errors.

In a statistical hypothesis testing experiment, a Type I error is committed by rejecting the null hypothesis when it is true. On the other hand, a Type II error is committed by not rejecting (accepting) the null hypothesis when it is false.

If we write ,

$$\alpha = P\ (\text{Type I error}) \quad = P\ (\text{rejecting } H_0 \mid H_0 \text{ is true})$$

$$\beta = P\ (\text{Type II error}) \quad = P\ (\text{Not rejecting } H_0 \mid H_0 \text{ is false})$$

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad. Thus we find ourselves in the situation which is described in the following table.

	Accept H_0	Reject H_0
H_0 is true	Correct decision	Type I Error
H_0 is false	Type II error	Correct decision

4.8 Test Procedure :

Steps for testing hypothesis is given below. (for both large sample and small sample tests)

1. Null hypothesis : set up null hypothesis H_0.
2. Alternative Hypothesis: Set up alternative hypothesis H_1, which is complementry to H_0 which will indicate whether one tailed (right or left tailed) or two tailed test is to be applied.
3. Level of significance : Choose an appropriate level of significance (α), α is fixed in advance.
4. Test statistic (or test of criterian):

   Calculate the value of the test statistic, Z $= \dfrac{t - E(t)}{S.E.(t)}$ under the null hypothesis, where t is the sample statistic
5. Inference: We compare the computed value of Z (in absolute value) with the significant value (critical value) $Z\alpha/2$ (or $Z\alpha$). If $|Z| > Z\alpha$, we reject the null hypothesis H_0 at α % level of significance and if $|Z| \leq Z\alpha$, we accept H_0 at α % level of significance.

Note:

1. Large Sample: A sample is large when it consists of more than 30 items.
2. Small Sample: A sample is small when it consists of 30 or less than 30 items.

Exercise -4

I. Choose the best answers:

1. A measure characterizing a sample such as $\overline{x}$ or s is called
   (a). Population (b). Statistic (c).Universe (d).Mean
2. The standard error of the mean is
   (a). σ^2 (b). $\frac{\sigma}{n}$ (c). $\frac{\sigma}{\sqrt{n}}$ (d). $\frac{\sqrt{n}}{\sigma}$
3. The standard error of observed sample proportion "P" is
   (a). $\sqrt{\frac{P(1-Q)}{n}}$ (b). $\sqrt{\frac{PQ}{n}}$ (c). $\sqrt{\frac{(1-P)Q}{n}}$ (d). $\frac{PQ}{n}$
4. Alternative hypothesis is
   (a). Always Left Tailed (b). Always Right tailed
   (c). Always One Tailed (d). One Tailed or Two Tailed
5. Critical region is
   (a). Rejection Area (b). Acceptance Area
   (c). Probability (d). Test Statistic Value
6. The critical value of the test statistic at level of significance α for a two tailed test is denoted by
   (a). $Z_{\alpha/2}$ (b).Z_{α} (c). $Z_{2\alpha}$ (d). $Z_{\alpha/4}$
7. In the right tailed test, the critical region is
   (a). 0 (b). 1
   (c). Lies entirely in right tail (d). Lies in the left tail
8. Critical value of $|Z_{\alpha}|$ at 5% level of significance for two tailed test is
   (a). 1.645 (b). 2.33 (c). 2.58 (d). 1.96

9. Under null hypothesis the value of the test statistic Z is

(a). $\frac{t - S.E.(t)}{E(t)}$ (b). $\frac{t + E(t)}{S.E.(t)}$ (c). $\frac{t - E(t)}{S.E.(t)}$ (d). $\sqrt{\frac{PQ}{n}}$

10. The alternative hypothesis H_1: $\mu \neq \mu_0$ ($\mu > \mu_0$ or $\mu < \mu_0$) takes the critical region as

(a). Right tail only (b). Both right and left tail
(c). Left tail only (d). Acceptance region

11. A hypothesis may be classified as

(a). Simple (b). Composite
(c). Null (d). All the above

12. Whether a test is one sided or two sided depends on

(a). Alternative hypothesis (b). Composite hypothesis
(c). Null hypothesis (d). Simple hypothesis

13. A wrong decision about H_0 leads to:

(a). One kind of error (b). Two kinds of error
(c). Three kinds of error (d). Four kinds of error

14. Area of the critical region depends on

(a). Size of type I error (b). Size of type II error
(c). Value of the statistics (d). Number of observations

15. Test of hypothesis $H_0 : \mu = 70$ vs $H_1 = \mu > 70$ leads to

(a). One sided left tailed test (b). One sided right tailed test
(c). Two tailed test (d). None of the above

16. Testing $H_0 : \mu = 1500$ against $\mu < 1500$ leads to

(a). One sided left tailed test (b). One sided right tailed test
(c). Two tailed test (d). All the above

17. Testing $H_0 : \mu = 100$ vs H_1: $\mu \neq 100$ lead to

(a). One sided right tailed test (b). One sided left tailed test
(c). Two tailed test (d). None of the above

II. Fill in the Blanks

18. n_1 and n_2 represent the __________ of the two independent random samples respectively.

19. Standard error of the observed sample proportion p is ________

20. When the hypothesis is true and the test rejects it, this is called ________

21. When the hypothesis is false and the test accepts it this is called _______

22. Formula to calculate the value of the statistic is __________

III. Answer the following

23. Define sampling distribution.
24. Define Parameter and Statistic.
25. Define standard error.
26. Give the standard error of the difference of two sample proportions.
27. Define Null hypothesis and alternative hypothesis.
28. Explain: Critical Value.
29. What do you mean by level of significance?
30. Explain clearly type I and type II errors.
31. What are the procedure generally followed in testing of a hypothesis ?
32. What do you mean by testing of hypothesis?
33. Write a detailed note on one- tailed and two-tailed tests.

Answers:

I.

1. (b) 2. (c) 3. (b) 4.(d) 5. (a)

6. (a) 7. (c) 8. (d) 9. (c) 10 (b)

11.(d) 12.(a) 13.(b) 14.(a) 15.(b)

16.(a) 17.(c)

II.

18. size 19. $\sqrt{\frac{PQ}{n}}$ 20. Type I error

21. Type II error

22. $Z = \frac{t - E(t)}{S.E.(t)}$

5. TEST OF SIGNIFICANCE (Large Sample)

5.0 Introduction:

In practical problems, statisticians are supposed to make tentative calculations based on sample observations. For example

(i) The average weight of school student is 35kg

(ii) The coin is unbiased

Now to reach such decisions it is essential to make certain assumptions (or guesses) about a population parameter. Such an assumption is known as statistical hypothesis, the validity of which is to be tested by analysing the sample. The procedure, which decides a certain hypothesis is true or false, is called the test of hypothesis (or test of significance).

Let us assume a certain value for a population mean. To test the validity of our assumption, we collect sample data and determine the difference between the hypothesized value and the actual value of the sample mean. Then, we judge whether the difference is significant or not. The smaller the difference, the greater the likelihood that our hypothesized value for the mean is correct. The larger the difference the smaller the likelihood, which our hypothesized value for the mean, is not correct.

5.1 Large samples ($n > 30$):

The tests of significance used for problems of large samples are different from those used in case of small samples as the assumptions used in both cases are different. The following assumptions are made for problems dealing with large samples:

(i) Almost all the sampling distributions follow normal asymptotically.

(ii) The sample values are approximately close to the population values.

The following tests are discussed in large sample tests.

(i) Test of significance for proportion

(ii) Test of significance for difference between two proportions

(iii) Test of significance for mean

(iv) Test of significance for difference between two means.

5.2 Test of Significance for Proportion:

Test Procedure

Set up the null and alternative hypotheses

$H_0 : P = P_0$

$H_1 = P \neq P_0$ $(P>P_0$ or $P<P_0)$

Level of significance:

Let $\alpha = 0\ .05$ or 0.01

Calculation of statistic:

Under H_0 the test statistic is

$$Z_0 = \left| \frac{p-P}{\sqrt{\frac{PQ}{n}}} \right|$$

Expected value:

$$Z_e = \left| \frac{p-P}{\sqrt{\frac{PQ}{n}}} \right| \sim N(0, 1)$$

$$= 1.96 \quad \text{for } \alpha = 0.05\ (1.645)$$

$$= 2.58 \quad \text{for } \alpha = 0.01\ \ (2.33)$$

Inference:

(i) If the computed value of $Z_0 \leq Z_e$ we accept the null hypothesis and conclude that the sample is drawn from the population with proportion of success P_0

(ii) If $Z_0 > Z_e$ we reject the null hypothesis and conclude that the sample has not been taken from the population whose population proportion of success is P_0.

Example 1:

In a random sample of 400 persons from a large population 120 are females.Can it be said that males and females are in the ratio 5:3 in the population? Use 1% level of significance

Solution:

We are given

n = 400 and

x = No. of female in the sample = 120

p = observed proportion of females in the sample = $\frac{120}{400}$ = 0.30

Null hypothesis:

The males and females in the population are in the ratio 5:3

i.e., H_0: P = Proportion of females in the population = $\frac{3}{8}$ = 0.375

Alternative Hypothesis:

$H_1 : P \neq 0.375$ (two-tailed)

Level of significance:

α = 1 % or 0.01

Calculation of statistic:

$$\text{Under } H_0\text{, the test statistic is } Z_0 = \left| \frac{p - P}{\sqrt{\frac{PQ}{n}}} \right|$$

$$= \left| \frac{0.300 - 0.375}{\sqrt{\frac{0.375 \times 0.625}{400}}} \right|$$

$$= \frac{0.075}{\sqrt{0.000586}} = \frac{0.075}{0.024} = 3.125$$

Expected value:

$$Z_e = \left| \frac{p - P}{\sqrt{\frac{PQ}{n}}} \right| \sim N(0,1) = 2.58$$

Inference :

Since the calculated $Z_0 > Z_e$ we reject our null hypothesis at 1% level of significance and conclude that the males and females in the population are not in the ratio 5:3

Example 2:

In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company, however, claimed that only 5% of their product is defective. Is the claim tenable?

Solution:

We are given n = 400

x = No. of defectives in the sample = 30

p= proportion of defectives in the sample

$$= \frac{x}{n} = \frac{30}{400} = 0.075$$

Null hypothesis:

The claim of the company is tenable H_0: P= 0.05

Alternative Hypothesis:

H_1 : P > 0.05 (Right tailed Alternative)

Level of significance: 5%

Calculation of statistic:

Under H_0, the test statistic is

$$Z_0 = \left|\frac{p - P}{\sqrt{\frac{PQ}{n}}}\right|$$

$$= \left|\frac{0.075 - 0.050}{\sqrt{\frac{0.05 \times 0.95}{400}}}\right|$$

$$= \frac{0.025}{\sqrt{0.0001187}} = 2.27$$

Expected value:

$$Z_e = \left| \frac{p - P}{\sqrt{\frac{PQ}{n}}} \right| \sim N(0, 1)$$

$$= 1.645 \text{ (Single tailed)}$$

Inference :

Since the calculated $Z_0 > Z_e$ we reject our null hypothesis at 5% level of significance and we conclude that the company' s claim is not tenable.

5.3 Test of significance for difference between two proportion:

Test Procedure

Set up the null and alternative hypotheses:

$H_0 : P_1 = P_2 = P$ (say)

$H_1 : P_1 \neq P_2$ ($P_1 > P_2$ or $P_1 < P_2$)

Level of significance:

Let $\alpha = 0.05$ or 0.01

Calculation of statistic:

Under H_0, the test statistic is

$$Z_0 = \left| \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \right| \quad \text{(}P_1 \text{ and } P_2 \text{ are known)}$$

$$= \left| \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \quad \text{(}P_1 \text{ and } P_2 \text{ are not known)}$$

where $\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$

$\hat{Q} = 1 - \hat{P}$

Expected value:

$$Z_e = \left|\frac{p_1 - p_2}{S.E(p_1 - p_2)}\right| \sim N(0,1)$$

Inference:

(i) If $Z_0 \leq Z_e$ we accept the null hypothesis and conclude that the difference between proportions are due to sampling fluctuations.

(ii) If $Z_0 > Z_e$ we reject the null hypothesis and conclude that the difference between proportions cannot be due to sampling fluctuations

Example 3:

In a referendum submitted to the ' student body' at a university, 850 men and 550 women voted. 530 of the men and 310 of the women voted ' yes' . Does this indicate a significant difference of the opinion on the matter between men and women students?

Solution:

We are given

n_1= 850 n_2 = 550 x_1= 530 x_2=310

$p_1 = \frac{530}{850} = 0.62$ $p_2 = \frac{310}{550} = 0.56$

$$\hat{P} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{530 + 310}{1400} = 0.60$$

$$\hat{Q} = 0.40$$

Null hypothesis:

H_0: P_1= P_2 ie the data does not indicate a significant difference of the opinion on the matter between men and women students.

Alternative Hypothesis:

$H_1 : P_1 \neq P_2$ (Two tailed Alternative)

Level of significance:

Let $\alpha = 0.05$

Calculation of statistic:

Under H_0, the test statistic is

$$Z_0 = \left| \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$

$$= \left| \frac{0.62 - 0.56}{\sqrt{0.6 \times 0.4\left(\frac{1}{850} + \frac{1}{550}\right)}} \right|$$

$$= \frac{0.06}{0.027} = 2.22$$

Expected value:

$$Z_e = \left| \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \sim N(0,1) = 1.96$$

Inference :

Since $Z_0 > Z_e$ we reject our null hypothesis at 5% level of significance and say that the data indicate a significant difference of the opinion on the matter between men and women students.

Example 4:

In a certain city 125 men in a sample of 500 are found to be self employed. In another city, the number of self employed are 375 in a random sample of 1000. Does this indicate that there is a greater population of self employed in the second city than in the first?

Solution:

We are given

n_1= 500 n_2 = 1000 x_1 = 125 x_2 = 375

$$p_1 = \frac{125}{500} = 0.25 \qquad p_2 = \frac{375}{1000} = 0.375$$

$$\hat{P} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{125+375}{500+1000}$$

$$= \frac{500}{1500} = \frac{1}{3}$$

$$\hat{Q} = 1 - \frac{1}{3} = \frac{2}{3}$$

Null hypothesis:

H_0: $P_1 = P_2$ There is no significant difference between the two population proportions.

Alternative Hypothesis:

$H_1 : P_1 < P_2$ (left tailed Alternative)

Level of significance: Let $\alpha = 0.05$

Calculation of statistic:

Under H_0, the test statistic is

$$Z_0 = \left| \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$

$$= \left| \frac{0.25 - 0.375}{\sqrt{\frac{1}{3} \times \frac{2}{3}\left(\frac{1}{500} + \frac{1}{1000}\right)}} \right| = \frac{0.125}{0.026} = 4.8$$

Expected value:

$$Z_e = \left| \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \sim N(0,1) = 1.645$$

Inference :

Since $Z_0 > Z_e$ we reject the null hypothesis at 5% level of significance and say that there is a significant difference between the two population proportions.

Example 5:

A civil service examination was given to 200 people. On the basis of their total scores, they were divided into the upper 30% and the remaining 70%. On a certain question 40 of the upper group and 80 of the lower group answered correctly. On the basis of this question, is this question likely to be useful for discriminating the ability of the type being tested?

Solution:

We are given

$n_1 = \frac{30 \times 200}{100} = 60$ $\qquad$ $n_2 = \frac{70 \times 200}{100} = 140$

$x_1 = 40$ $\qquad$ $x_2 = 80$

$p_1 = \frac{40}{60} = \frac{2}{3}$ $\qquad$ $p_2 = \frac{80}{140} = \frac{4}{7}$

$$\hat{P} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{40 + 80}{60 + 140}$$

$$= \frac{120}{200} = \frac{6}{10}$$

$$\hat{Q} = 1 - \hat{P} = 1 - \frac{1}{6} = \frac{4}{10}$$

Null hypothesis:

H_0: $P_1 = P_2$ (say) The particular question does not discriminate the abilities of two groups.

Alternative Hypothesis:

H_1 : $P_1 \neq P_2$ (two tailed Alternative)

Level of significance:

Let $\alpha = 0.05$

Calculation of statistics

Under H_0, the test statistic is

$$Z_0 = \left| \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$

$$= \left| \frac{\frac{2}{3} - \frac{4}{7}}{\sqrt{\frac{6}{10} \times \frac{4}{10}\left(\frac{1}{60} + \frac{1}{140}\right)}} \right|$$

$$= \frac{10}{\sqrt{21}\sqrt{3}} = 1.3$$

Expected value:

$$Z_e = \left| \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \sim \quad N(0,1)$$

$$= 1.96 \text{ for } \alpha=0.05$$

Inference :

Since $Z_0 < Z_e$ we accept our null hypothesis at 5% level of significance and say that the particular question does not discriminate the abilities of two groups.

5.4 Test of significance for mean:

Let x_i (i = 1,2….n) be a random sample of size n from a population with variance σ^2, then the sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots x_n)$$

$$E(\bar{x}) = \mu$$

$$V(\bar{x}) = V\left[\frac{1}{n}(x_1 + x_2 + \ldots x_n)\right]$$

$$= \frac{1}{n^2}[(V(x_1) + V(x_2) + \ldots V(x_n)]$$

$$= \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}$$

$$\therefore \text{ S.E } (\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Test Procedure:

Null and Alternative Hypotheses:

$H_0:\mu = \mu_0$.

$H_1:\mu \neq \mu_0$ ($\mu > \mu_0$ or $\mu < \mu_0$)

Level of significance:

Let $\alpha = 0.05$ or 0.01

Calculation of statistic:

Under H_0, the test statistic is

$$Z_0 = \left|\frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})}\right| = \left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right|$$

Expected value:

$$Z_e = \left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \sim N(0,1)$$

$= 1.96$ for $\alpha = 0.05$ (1.645)

or

$= 2.58$ for $\alpha = 0.01$ (2.33)

Inference :

If $Z_0 \leq Z_{e,}$ we accept our null hypothesis and conclude that the sample is drawn from a population with mean $\mu = \mu_0$

If $Z_0 > Z_e$ we reject our H_0 and conclude that the sample is not drawn from a population with mean $\mu = \mu_0$

Example 6:

The mean lifetime of 100 fluorescent light bulbs produced by a company is computed to be 1570 hours with a standard deviation of 120 hours. If μ is the mean lifetime of all the bulbs produced by the company, test the hypothesis μ=1600 hours against the alternative hypothesis $\mu \neq 1600$ hours using a 5% level of significance.

Solution:

We are given

$\bar{x} = 1570$ hrs $\mu = 1600$hrs $s = 120$ hrs $n=100$

Null hypothesis:

H_0: $\mu = 1600$.ie There is no significant difference between the sample mean and population mean.

Alternative Hypothesis:

H_1: $\mu \neq 1600$ (two tailed Alternative)

Level of significance:

Let $\alpha = 0.05$

Calculation of statistics

Under H_0, the test statistic is

$$Z_0 = \left|\frac{\bar{x}-\mu}{s/\sqrt{n}}\right|$$

$$= \left|\frac{1570-1600}{\frac{120}{\sqrt{100}}}\right|$$

$$= \frac{30\times10}{120}$$

$$= 2.5$$

Expected value:

$$Z_0 = \left|\frac{\bar{x}-\mu}{s/\sqrt{n}}\right| \sim N(0,1)$$

$$= 1.96 \text{ for } \alpha = 0.05$$

Inference :

Since $Z_0 > Z_e$ we reject our null hypothesis at 5% level of significance and say that there is significant difference between the sample mean and the population mean.

Example 7:

A car company decided to introduce a new car whose mean petrol consumption is claimed to be lower than that of the existing car. A sample of 50 new cars were taken and tested for petrol consumption. It was found that mean petrol consumption for the 50 cars was 30 km per litre with a standard deviation of 3.5 km per litre. Test at 5% level of significance whether the company's claim that the new car petrol consumption is 28 km per litre on the average is acceptable.

Solution:

We are given $\overline{x} = 30$; $\mu = 28$; $n=50$; $s=3.5$

Null hypothesis:

H_0: $\mu = 28$. i.e The company's claim that the petrol consumption of new car is 28km per litre on the average is acceptable.

Alternative Hypothesis:

H_1: $\mu < 28$ (Left tailed Alternative)

Level of significance:

Let $\alpha = 0.05$

Calculation of statistic:

Under H_0 the test statistics is

$$Z_0 = \left|\frac{\overline{x}-\mu}{s/\sqrt{n}}\right|$$

$$= \left|\frac{30-28}{\frac{3.5}{\sqrt{50}}}\right|$$

$$= \frac{2\times\sqrt{50}}{3.5}$$

$$= 4.04$$

Expected value:

$$Z_e = \left|\frac{\overline{x}-\mu}{s/\sqrt{n}}\right| \sim N(0,1) \text{ at } \alpha = 0.05$$

$$= 1.645$$

Inference :

Since the calculated $Z_0 > Z_e$ we reject the null hypothesis at 5% level of significance and conclude that the company' s claim is not acceptable.

5.5 Test of significance for difference between two means:

Test procedure

Set up the null and alternative hypothesis

H_0: $\mu_1 = \mu_2$; **H_1:** $\mu_1 \neq \mu_2$ ($\mu_1 > \mu_2$ or $\mu_1 < \mu_2$)

Level of significance:

Let α%

Calculation of statistic:

Under H_0 the test statistic is

$$Z_0 = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^{\;2}}{n_1} + \frac{\sigma_2^{\;2}}{n_2}}} \right|$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (ie) If the samples have been drawn from the population with common S.D σ then under $H_0 : \mu_1 = \mu_2$

$$Z_0 = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|$$

Expected value:

$$Z_e = \left| \frac{\overline{x}_1 - \overline{x}_2}{S.E(\overline{x}_1 - \overline{x}_2)} \right| \sim N(0,1)$$

Inference:

(i) If $Z_0 \leq Z_e$ we accept the H_0 (ii) If $Z_0 > Z_e$ we reject the H_0

Example 8:

A test of the breaking strengths of two different types of cables was conducted using samples of $n_1 = n_2 = 100$ pieces of each type of cable.

Cable I	Cable II
$\overline{x}_1$ =1925	$\overline{x}_2$ = 1905
σ_1= 40	σ_2 = 30

Do the data provide sufficient evidence to indicate a difference between the mean breaking strengths of the two cables? Use 0.01 level of significance.

Solution:

We are given

$\overline{x}_1$ =1925 $\overline{x}_2$ = 1905 σ_1= 40 σ_2 = 30

Null hypothesis

$H_0:\mu_1 = \mu_2$ .ie There is no significant difference between the mean breaking strengths of the two cables.

$H_1 : \mu_1 \neq \mu_2$ (Two tailed alternative)

Level of significance:

Let α = 0.01 or 1%

Calculation of statistic:

Under H_0 the test statistic is

$$Z_0 = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right|$$

$$= \left| \frac{1925 - 1905}{\sqrt{\frac{40^2}{100} + \frac{30^2}{100}}} \right| = \frac{20}{5} = 4$$

Expected value:

$$Z_e = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| \sim N(0,1) = 2.58$$

Inference:

Since $Z_0 > Z_{e,}$ we reject the $H_{0.}$ Hence the formulated null hypothesis is wrong ie there is a significant difference in the breaking strengths of two cables.

Example 9:

The means of two large samples of 1000 and 2000 items are 67.5 cms and 68.0cms respectively. Can the samples be regarded as drawn from the population with standard deviation 2.5 cms. Test at 5% level of significance.

Solution:

We are given

$n_1 = 1000$; $n_2 = 2000$ $\overline{x}_1 = 67.5$ cms ; $\overline{x}_2 = 68.0$ cms $\sigma = 2.5$ cms

Null hypothesis

H_0: $\mu_1 = \mu_2$ (i.e.,) the sample have been drawn from the same population.

Alternative Hypothesis:

H_1: $\mu_1 \neq \mu_2$ (Two tailed alternative)

Level of significance:

$\alpha = 5\%$

Calculation of statistic:

Under H_o the test statistic is

$$Z_0 = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|$$

$$= \left| \frac{67.5 - 68.0}{2.5\sqrt{\frac{1}{1000} + \frac{1}{2000}}} \right|$$

$$= \frac{0.5 \times 20}{2.5\sqrt{3/5}}$$

$= 5.1$

Expected value:

$$Z_e = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \sim N(0,1) = 1.96$$

Inference :

Since $Z_0 > Z_e$ we reject the H_0 at 5% level of significance and conclude that the samples have not come from the same population.

Exercise – 5

I. Choose the best answer:

1. Standard error of number of success is given by

   (a) $\sqrt{\frac{pq}{n}}$ (b) $\sqrt{npq}$ (c) npq (d) $\sqrt{\frac{np}{q}}$

2. Large sample theory is applicable when

   (a) $n > 30$ (b) $n < 30$ (c) $n < 100$ (d) $n < 1000$

3. Test statistic for difference between two means is

   (a) $\frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$ (b) $\frac{p - P}{\sqrt{\frac{PQ}{n}}}$

   (c) $\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ (d) $\frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

4. Standard error of the difference of proportions (p_1-p_2) in two classes under the hypothesis H_0:$p_1 = p_2$ with usual notation is

   (a) $\sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}$ (b) $\sqrt{\hat{p}(\frac{1}{n_1} + \frac{1}{n_2})}$

   (c) $\hat{p}\hat{q}\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}$ (d) $\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

5. Statistic z = $\dfrac{\bar{x}-\bar{y}}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ is used to test the null hypothesis

(a) $H_{0:}\ \mu_1+\mu_2=0$ (b) $H_{0:}\mu_1-\mu_2=0$

(c) H_0: $\mu=\mu_0$(a constant) (c) none of the above.

II. Fill in the blanks:

6. If $\hat{P}=\frac{2}{3}$, then $\hat{Q}=$ ________
7. If $z_0 < z_e$ then the null hypothesis is ____________
8. When the difference is __________, the null hypothesis is rejected.
9. Test statistic for difference between two proportions is ________
10. The variance of sample mean is ________

III. Answer the following

11. In a test if $z_0 \le z_e$, what is your conclusion about the null hypothesis?
12. Give the test statistic for
    (a) Proportion
    (b) Mean
    (c) Difference between two means
    (d) Difference between two proportions
13. Write the variance of difference between two proportions
14. Write the standard error of proportion.
15. Write the test procedure for testing the test of significance for
    (a) Proportion (b) mean
    (c) difference between two proportions
    (d) difference between two mean
16. A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.
17. A person throws 10 dice 500 times and obtains 2560 times 4, 5 or 6. Can this be attributed to fluctuations of sampling?

18. In a hospital 480 female and 520 male babies were born in a week. Do these figure confirm the hypothesis that males and females are born in equal number?
19. In a big city 325 men out of 600 men were found to be self-employed. Does this information support the conclusion that the majority of men in this city are self-employed?
20. A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?
21. In a random sample of 1000 persons from town A , 400 are found to be consumers of wheat. In a sample of 800 from town B, 400 are found to be consumers of wheat. Do these data reveal a significant difference between town A and town B, so far as the proportion of wheat consumers is concerned?
22. 1000 articles from a factory A are examined and found to have 3% defectives. 1500 similar articles from a second factory B are found to have only 2% defectives. Can it be reasonably concluded that the product of the first factory is inferior to the second?
23. In a sample of 600 students of a certain college, 400 are found to use blue ink. In another college from a sample of 900 students 450 are found to use blue ink. Test whether the two colleges are significantly different with respect to the habit of using blue ink.
24. It is claimed that a random sample of 100 tyres with a mean life of 15269kms is drawn from a population of tyres which has a mean life of 15200 kms and a standard deviation of 1248 kms. Test the validity of the claim.
25. A sample of size 400 was drawn and the sample mean B was found to be 99. Test whether this sample could have come from a normal population with mean 100 and variance 64 at 5% level of significance.
26. The arithmetic mean of a sample 100 items drawn from a large population is 52. If the standard deviation of the population is 7, test the hypothesis that the mean of the population is 55 against the alternative that the mean is not 55. Use 5% level of significance.

27. A company producing light bulbs finds that mean life span of the population of bulbs is 1200 hrs with a standard deviation of 125hrs. A sample of 100 bulbs produced in a lot is found to have a mean life span of 1150hrs. Test whether the difference between the population and sample means is statistically significant at 5% level of significance.

28. Test the significance of the difference between the means of the samples from the following data

	Size of sample	Mean	Standard deviation
Sample A	100	50	4
Sample B	150	51	5

29. An examination was given to two classes consisting of 40 and 50 students respectively. In the first class the mean mark was 74 with a standard deviation of 8, while in the second class the mean mark was 78 with a standard deviation of 7. Is there a significant difference between the performance of the two classes at a level of significance of 0.05?

30. If 60 M.A. Economics students are found to have a mean height of 63.60 inches and 50 M.Com students a mean height of 69.51 inches. Would you conclude that the commerce students are taller than Economics students? Assume the standard deviation of height of post-graduate students to be 2.48 inches.

Answers:

I.

1. (b) 2.(a) 3.(c) 4.(a) 5.(b)

II.

6. $\frac{1}{3}$ 7.accepted 8.significant

9. $\dfrac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$ 10. $\dfrac{\sigma^2}{n}$

III.

16. z = 1.6 Accept H_0 17. z = 1.7 Accept H_0

18. z = 1.265Accept H_0 19. z = 2.04 Accept H_0

20. z = 0.106 Accept H_0 21. z = 4.247 Reject H_0

22. z = 1.63 Accept H_0 23. z = 6.38 Reject H_0

24. z = 0.5529 Accept H_0 25. z = 2.5 Reject H_0

26. z = 4.29 Reject H_0 27. z = 4 Reject H_0

28. z = 1.75 Accept H_0 29. z = 2.49 Reject H_0

30. z = 12.49 Reject H_0

6. TESTS OF SIGNIFICANCE
(Small Samples)

6.0 Introduction:

In the previous chapter we have discussed problems relating to large samples. The large sampling theory is based upon two important assumptions such as

(a) The random sampling distribution of a statistic is approximately normal and

(b) The values given by the sample data are sufficiently close to the population values and can be used in their place for the calculation of the standard error of the estimate.

The above assumptions do not hold good in the theory of small samples. Thus, a new technique is needed to deal with the theory of small samples. A sample is small when it consists of less than 30 items. ($n< 30$)

Since in many of the problems it becomes necessary to take a small size sample, considerable attention has been paid in developing suitable tests for dealing with problems of small samples. The greatest contribution to the theory of small samples is that of Sir William Gosset and Prof. R.A. Fisher. Sir William Gosset published his discovery in 1905 under the pen name ‘ Student’ and later on developed and extended by Prof. R.A.Fisher. He gave a test popularly known as ‘ t-test’ .

6.1 t - statistic definition:

If x_1, x_2, ….x_n is a random sample of size n from a normal population with mean μ and variance σ^2, then Student’ s t-statistic is defined as $t = \dfrac{\bar{x}-\mu}{\dfrac{S}{\sqrt{n}}}$

where $\bar{x} = \dfrac{\sum x}{n}$ is the sample mean and $S^2 = \frac{1}{n-1}\sum(x-\overline{x})^2$

is an unbiased estimate of the population variance σ^2 It follows student' s t-distribution with $\nu = n-1$ d.f

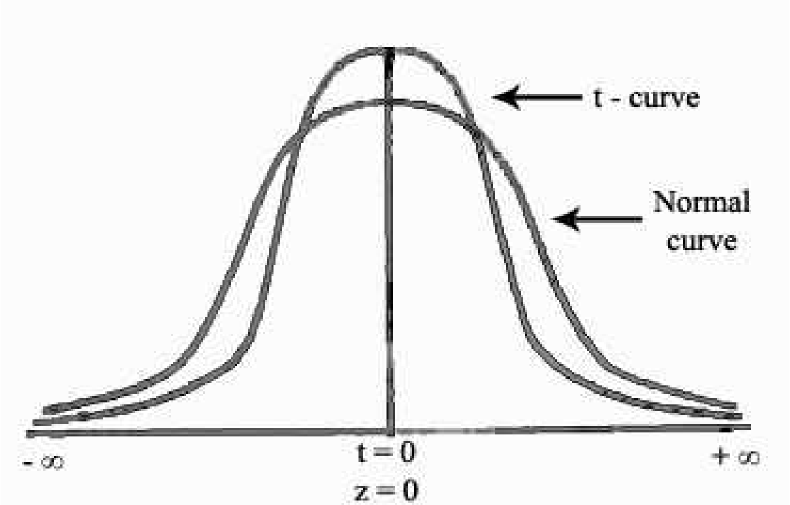
6.1.1 Assumptions for students t-test:

1. The parent population from which the sample drawn is normal.
2. The sample observations are random and independent.
3. The population standard deviation σ is not known.

6.1.2 Properties of t- distribution:

1. t-distribution ranges from $-\infty$ to ∞ just as does a normal distribution.
2. Like the normal distribution, t-distribution also symmetrical and has a mean zero.
3. t-distribution has a greater dispersion than the standard normal distribution.
4. As the sample size approaches 30, the t-distribution, approaches the Normal distribution.

Comparison between Normal curve and corresponding t - curve:



6.1.3 Degrees of freedom (d.f):

Suppose it is asked to write any four number then one will have all the numbers of his choice. If a restriction is applied or imposed to the choice that the sum of these number should be 50. Here, we have a choice to select any three numbers, say 10, 15, 20 and the fourth number is 5: [50 – (10 +15+20)]. Thus our choice of freedom is reduced by one, on the condition that the total be 50. therefore the restriction placed on the freedom is one and degree of freedom is three. As the restrictions increase, the freedom is reduced.

The number of independent variates which make up the statistic is known as the degrees of freedom and is usually denoted by ν (Nu)

The number of degrees of freedom for n observations is n – k where k is the number of independent linear constraint imposed upon them.

For the student' s t-distribution. The number of degrees of freedom is the sample size minus one. It is denoted by $\nu = n - 1$

The degrees of freedom plays a very important role in χ^2 test of a hypothesis.

When we fit a distribution the number of degrees of freedom is (n– k–1) where n is number of observations and k is number of parameters estimated from the data.

For e.g., when we fit a Poisson distribution the degrees of freedom is $\nu = n - 1 - 1$

In a contingency table the degrees of freedom is (r–1) (c –1) where r refers to number rows and c refers to number of columns.

Thus in a 3 × 4 table the d.f are (3–1) (4–1) = 6 d.f In a 2 × 2 contingency table the d.f are (2–1) (2–1) = 1

In case of data that are given in the form of series of variables in a row or column the d.f will be the number of observations in a series less one ie $\nu = n - 1$

Critical value of t:

The column figures in the main body of the table come under the headings $t_{0.100}$, $t_{0.50}$, $t_{0.025}$, $t_{0.010}$ and $t_{0.005}$. The subscripts give the proportion of the distribution in ‘tail’ area. Thus for two-tailed test at 5% level of significance there will be two rejection areas each containing 2.5% of the total area and the required column is headed $t_{0.025}$

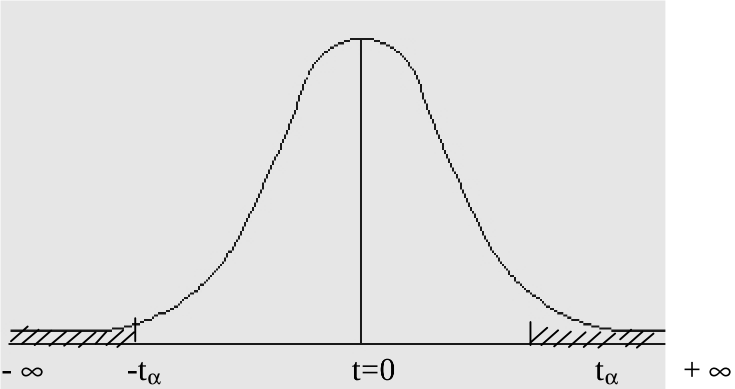
For example,

t_v (.05) for single tailed test = t_v (0.025) for two tailed test

t_v (.01) for single tailed test = t_v (0.005) for two tailed test

Thus for one tailed test at 5% level the rejection area lies in one end of the tail of the distribution and the required column is headed $t_{0.05}$.

Critical value of t – distribution



6.1.4 Applications of t-distribution:

The t-distribution has a number of applications in statistics, of which we shall discuss the following in the coming sections:

(i) t-test for significance of single mean, population variance being unknown.

(ii) t-test for significance of the difference between two sample means, the population variances being equal but unknown.

(a) Independent samples

(b) Related samples: paired t-test

6.2 Test of significance for Mean:

We set up the corresponding null and alternative hypotheses as follows:

H_0: $\mu = \mu_0$; There is no significant difference between the sample mean and population Mean.

H_1: $\mu \neq \mu_0$ ($\mu < \mu_0$ (or) $\mu > \mu_0$)

Level of significance:

5% or 1%

Calculation of statistic:

Under H_0 the test statistic is

$$t_0 = \left| \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \right| \quad \text{or} \quad \left| \frac{\bar{x} - \mu}{s / \sqrt{n-1}} \right|$$

where $\bar{x} = \frac{\sum x}{n}$ is the sample mean

and $S^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$ (or) $s^2 = \frac{1}{n} \sum (x - \bar{x})^2$

Expected value :

$$t_e = \left| \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \right| \sim \text{student' s t-distribution with (n-1) d.f}$$

Inference :

If $t_0 \leq t_e$ it falls in the acceptance region and the null hypothesis is accepted and if $t_o > t_e$ the null hypothesis H_0 may be rejected at the given level of significance.

Example 1:

Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weigh (in kg) as follows:

50 49 52 44 45 48 46 45 49 45

Test if the average packing can be taken to be 50 kg.

Solution:

Null hypothesis:

$H_0 : \mu = 50$ kgs in the average packing is 50 kgs.

Alternative Hypothesis:

$H_1 : \mu \neq 50$kgs (Two -tailed)

Level of Significance:

Let $\alpha = 0.05$

Calculation of sample mean and S.D

X	d = x –48	d^2
50	2	4
49	1	1
52	4	16
44	–4	16
45	–3	9
48	0	0
46	–2	4
45	–3	9
49	+1	1
45	–3	9
Total	–7	69

$$\bar{x} = A + \frac{\sum d}{n}$$

$$= 48 + \frac{-7}{10}$$

$$= 48–0.7 =47.3$$

$$S^2 = \frac{1}{n-1}[\sum d^2 - \frac{(\sum d)^2}{n}]$$

$$= \frac{1}{9}[69 - \frac{(7^2)}{10}]$$

$$= \frac{64.1}{9} = 7.12$$

Calculation of Statistic:

Under H_0 the test statistic is :

$$t_0 = \left| \frac{\bar{x} - \mu}{\sqrt{S^2 / n}} \right|$$

$$= \left|\frac{47.3 - 50.0}{\sqrt{7.12/10}}\right|$$

$$= \frac{2.7}{\sqrt{0.712}} \quad = 3.2$$

Expected value:

$t_e = \left|\frac{\bar{x} - \mu}{\sqrt{S^2/n}}\right|$ follows t distribution with (10–1) d.f

$= 2.262$

Inference:

Since $t_0 > t_e$, H_0 is rejected at 5% level of significance and we conclude that the average packing cannot be taken to be 50 kgs.

Example 2:

A soap manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign, the mean sales per week per shop was 140 dozens. After the campaign, a sample of 26 shops was taken and the mean sales was found to be 147 dozens with standard deviation 16. Can you consider the advertisement effective?

Solution:

We are given

$n = 26;$ $\bar{x} = 147$dozens; $s = 16$

Null hypothesis:

H_0: $\mu = 140$ dozens i.e. Advertisement is not effective.

Alternative Hypothesis:

H_1: $\mu > 140$kgs (Right -tailed)

Calculation of statistic:

Under the null hypothesis H_0, the test statistic is

$$t_0 = \left|\frac{\bar{x} - \mu}{s/\sqrt{n-1}}\right|$$

$$= \left|\frac{147 - 140}{16/\sqrt{25}}\right| \quad = \frac{7 \times 5}{16} \quad = 2.19$$

Expected value:

$$t_e = \left| \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \right| \quad \text{follows t-distribution with (26-1) = 25d.f}$$

$$= 1.708$$

Inference:

Since $t_0 > t_{e,}$ H_0 is rejected at 5% level of significance. Hence we conclude that advertisement is certainly effective in increasing the sales.

6.3 Test of significance for difference between two means:

6.3.1 Independent samples:

Suppose we want to test if two independent samples have been drawn from two normal populations having the same means, the population variances being equal. Let $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2, \ldots y_{n_2}$ be two independent random samples from the given normal populations.

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ i.e. the samples have been drawn from the normal populations with same means.

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ $(\mu_1 < \mu_2$ or $\mu_1 > \mu_2)$

Test statistic:

Under the H_0, the test statistic is

$$t_0 = \left| \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|$$

$$\text{where} \quad \bar{x} = \frac{\Sigma x}{n_1} \; ; \; \bar{y} = \frac{\Sigma y}{n_2}$$

$$\text{and } S^2 = \frac{1}{n_1 + n_2 - 2} \; [\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2] = \frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}$$

Expected value:

$$t_e = \left| \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|$$ follows t-distribution with $n_1 + n_2 - 2$ d.f

Inference:

If the $t_0 < t_e$ we accept the null hypothesis. If $t_0 > t_e$ we reject the null hypothesis.

Example 3:

A group of 5 patients treated with medicine ' A' weigh 42, 39, 48, 60 and 41 kgs: Second group of 7 patients from the same hospital treated with medicine ' B' weigh 38, 42 , 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine ' B' increases the weight significantly?

Solution:

Let the weights (in kgs) of the patients treated with medicines A and B be denoted by variables X and Y respectively.

Null hypothesis:

$H_0 : \mu_1 = \mu_2$

i.e. There is no significant difference between the medicines A and B as regards their effect on increase in weight.

Alternative Hypothesis:

$H_1 : \mu_1 < \mu_2$ (left-tail) i.e. medicine B increases the weight significantly.

Level of significance : Let $\alpha = 0.05$

Computation of sample means and S.Ds

Medicine A

X	$x - \bar{x}$ ($\bar{x} = 46$)	$(x - \bar{x})^2$
42	– 4	16
39	–7	49
48	2	4
60	14	196
41	– 5	25
230	**0**	**290**

$$\bar{x} = \frac{\sum x}{n_1} = \frac{230}{5} = 46$$

Medicine B

Y	$y - \bar{y}$ $(\bar{y} = 57)$	$(y - \bar{y})^2$
38	–19	361
42	–15	225
56	–1	1
64	7	49
68	11	121
69	12	144
62	5	25
399	**0**	**926**

$$\bar{y} = \frac{\sum y}{n_2} = \frac{399}{7} = 57$$

$$S^2 = \frac{1}{n_1 + n_2 - 2} [\sum(x - \overline{x)}^2 + \sum(y - \bar{y})^2]$$

$$= \frac{1}{10} [\ 290 + 926] = 121.6$$

Calculation of statistic:

Under H_0 the test statistic is

$$t_0 = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$

$$= \left| \frac{46 - 57}{\sqrt{121.6\left(\frac{1}{5} + \frac{1}{7}\right)}} \right|$$

$$= \frac{11}{\sqrt{121.6 \times \frac{12}{35}}}$$

$$= \frac{11}{6.57} = 1.7$$

Expected value:

$$t_e = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$ follows t-distribution with (5+7–2) =10 d.f

$= 1.812$

Inference:

Since $t_0 < t_e$ it is not significant. Hence H_0 is accepted and we conclude that the medicines A and B do not differ significantly as regards their effect on increase in weight.

Example 4:

Two types of batteries are tested for their length of life and the following data are obtained:

	No of samples	Mean life (in hrs)	Variance
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means?

Solution:

We are given

$n_1=9$; $\bar{x}_1=600$hrs; $s_1^2 =121$; $n_2 =8$; $\bar{x}_2 =640$hrs; $s_2^2=144$

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ i.e. Two types of batteries A and B are identical i.e. there is no significant difference between two types of batteries.

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ (Two- tailed)

Level of Significance:

Let $\alpha = 5\%$

Calculation of statistics:

Under H_0, the test statistic is

$$t_0 = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$

$$\text{where } S^2 = \frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2}$$

$$= \frac{9 \times 121 + 8 \times 144}{9 + 8 - 2}$$

$$= \frac{2241}{15} = 149.4$$

$$\therefore \quad t_0 = \left| \frac{600 - 640}{\sqrt{149.4\left(\frac{1}{9} + \frac{1}{8}\right)}} \right|$$

$$= \frac{40}{\sqrt{149.4\left(\frac{17}{72}\right)}} = \frac{40}{5.9391} = 6.735$$

Expected value:

$$t_e = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \quad \text{follows t-distribution with } 9+8-2 = 15 \text{ d.f}$$

$= 2.131$

Inference:

Since $t_0 > t_e$ it is highly significant. Hence H_0 is rejected and we conclude that the two types of batteries differ significantly as regards their length of life.

6.3.2 Related samples –Paired t-test:

In the t-test for difference of means, the two samples were independent of each other. Let us now take a particular situations where

(i) The sample sizes are equal; i.e., $n_1 = n_2 = n$(say), and

(ii) The sample observations (x_1, x_2, ...x_n) and (y_1, y_2, ...y_n) are not completely independent but they are dependent in pairs.

That is we are making two observations one before treatment and another after the treatment on the same individual. For example a business concern wants to find if a particular media of promoting sales of a product, say door to door canvassing or advertisement in papers or through T.V. is really effective. Similarly a pharmaceutical company wants to test the efficiency of a particular drug, say for inducing sleep after the drug is given. For testing of such claims gives rise to situations in (i) and (ii) above, we apply paired t-test.

Paired – t –test:

Let di = Xi – Yi (i = 1, 2, ...n) denote the difference in the observations for the i^{th} unit.

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ ie the increments are just by chance

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ ($\mu_1 > \mu_2$ (or) $\mu_1 < \mu_2$)

Calculation of test statistic:

$$t_0 = \left| \frac{\bar{d}}{S/\sqrt{n}} \right|$$

where $\bar{d} = \frac{\sum d}{n}$ and $S^2 = \frac{1}{n-1}\sum(d-\bar{d})^2 = \frac{1}{n-1}[\sum d^2 - \frac{(\sum d)^2}{n}]$

Expected value:

$$t_e = \left| \frac{\bar{d}}{S/\sqrt{n}} \right| \text{ follows t-distribution with } n-1 \text{ d.f}$$

Inference:

By comparing t_0 and t_e at the desired level of significance, usually 5% or 1%, we reject or accept the null hypothesis.

Example 5:

To test the desirability of a certain modification in typists desks, 9 typists were given two tests of as nearly as possible the same nature, one on the desk in use and the other on the new type. The following difference in the number of words typed per minute were recorded:

Typists	A	B	C	D	E	F	G	H	I
Increase in number of words	2	4	0	3	–1	4	–3	2	5

Do the data indicate the modification in desk promotes speed in typing?

Solution:

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ i.e. the modification in desk does not promote speed in typing.

Alternative Hypothesis:

$H_1 : \mu_1 < \mu_2$ (Left tailed test)

Level of significance: Let $\alpha = 0.05$

Typist	d	d^2
A	2	4
B	4	16
C	0	0
D	3	9
E	–1	1
F	4	16
G	–3	9
H	2	4
I	5	25
	$\Sigma d = 16$	$\Sigma d^2 = 84$

$$\bar{d} = \frac{\sum d}{n} = \frac{16}{9} = 1.778$$

$$S = \sqrt{\frac{1}{n-1}[\sum d^2 - \frac{(\sum d)^2}{n}]}$$

$$= \sqrt{\frac{1}{8}[84 - \frac{(16)^2}{9}]} = \sqrt{6.9} = 2.635$$

Calculation of statistic:

Under H_0 the test statistic is

$$t_0 = \left|\frac{\bar{d}.\sqrt{n}}{S}\right| = \frac{1.778 \times 3}{2.635} = 2.024$$

Expected value:

$t_e = \left|\frac{\bar{d}.\sqrt{n}}{S}\right|$ follows t- distribution with 9 –1 = 8 d.f

= 1.860

Inference:

When $t_0 < t_e$ the null hypothesis is accepted. The data does not indicate that the modification in desk promotes speed in typing.

Example 6:

An IQ test was administered to 5 persons before and after they were trained. The results are given below:

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

Test whether there is any change in IQ after the training programme (test at 1% level of significance)

Solution:

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ i.e. there is no significant change in IQ after the training programme.

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ (two tailed test)

Level of significance :

$\alpha = 0.01$

x	110	120	123	132	125	Total
y	120	118	125	136	121	-
d = x–y	–10	2	–2	–4	4	–10
d^2	100	4	4	16	16	140

$$\bar{d} = \frac{\sum d}{n} = \frac{-10}{5} = -2$$

$$S^2 = \frac{1}{n-1}[\sum d^2 - \frac{(\sum d)^2}{n}]$$

$$= \frac{1}{4}[140 - \frac{100}{5}] = 30$$

Calculation of Statistic:

Under H_0 the test statistic is

$$t_0 = \left|\frac{\bar{d}}{S/\sqrt{n}}\right|$$

$$= \left|\frac{-2}{\sqrt{30/5}}\right|$$

$$= \frac{2}{2.45}$$

$$= 0.816$$

Expected value:

$$t_e = \left|\frac{\bar{d}}{\sqrt{S^2/n}}\right| \quad \text{follows t-distribution with } 5-1 = 4 \text{ d.f}$$

$$= 4.604$$

Inference:

Since $t_0 < t_e$ at 1% level of significance we accept the null hypothesis. We therefore, conclude that there is no change in IQ after the training programme.

6.4 Chi square statistic:

Various tests of significance described previously have mostly applicable to only quantitative data and usually to the data which are approximately normally distributed. It may also happens that we may have data which are not normally distributed. Therefore there arises a need for other methods which are more appropriate for studying the differences between the expected and observed frequencies. The other method is called Non-parametric or distribution free test. A non- parametric test may be defined as a statistical test in which no hypothesis is made about specific values of parameters. Such non-parametric test has assumed great importance in statistical analysis because it is easy to compute.

6.4.1 Definition:

The Chi- square (χ^2) test (Chi-pronounced as ki) is one of the simplest and most widely used non-parametric tests in statistical work. The χ^2 test was first used by Karl Pearson in the year 1900. The quantity χ^2 describes the magnitude of the discrepancy between theory and observation. It is defined as

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(Oi - Ei)^2}{Ei}\right]$$

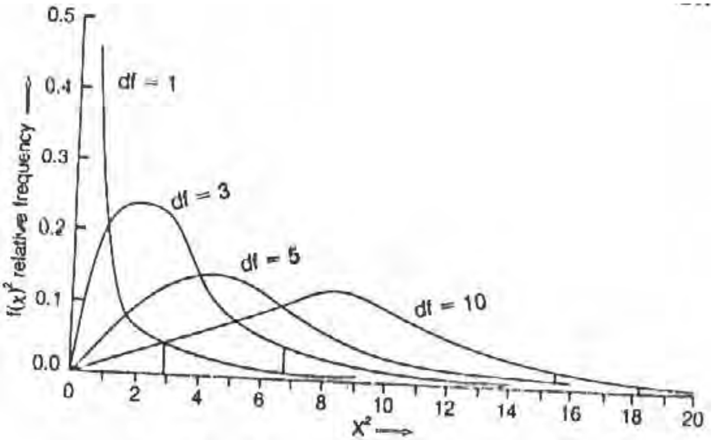
Where O refers to the observed frequencies and E refers to the expected frequencies.

Note:

If χ^2 is zero, it means that the observed and expected frequencies coincide with each other. The greater the discrepancy between the observed and expected frequencies the greater is the value of χ^2.

Chi square - Distribution:

The square of a standard normal variate is a Chi-square variate with 1 degree of freedom i.e., If X is normally distributed with mean μ and standard deviation σ, then $\left(\frac{x-\mu}{\sigma}\right)^2$ is a Chi-square variate (χ^2) with 1 d.f. The distribution of Chi-square depends on the degrees of freedom. There is a different distribution for each number of degrees of freedom.



6.4.2 properties of Chi-square distribution:

1. The Mean of χ^2 distribution is equal to the number of degrees of freedom (n)
2. The variance of χ^2 distribution is equal to 2n
3. The median of χ^2 distribution divides, the area of the curve into two equal parts, each part being 0.5.
4. The mode of χ^2 distribution is equal to (n–2)
5. Since Chi-square values always positive, the Chi square curve is always positively skewed.
6. Since Chi-square values increase with the increase in the degrees of freedom, there is a new Chi-square distribution with every increase in the number of degrees of freedom.
7. The lowest value of Chi-square is zero and the highest value is infinity ie $\chi^2 \geq 0$.
8. When Two Chi- squares χ_1^2 and χ_2^2 are independent χ^2 distribution with n_1 and n_2 degrees of freedom and their sum $\chi_1^2 + \chi_2^2$ will follow χ^2 distribution with $(n_1 + n_2)$ degrees of freedom.
9. When n (d.f) > 30, the distribution of $\sqrt{2\chi^2}$ approximately follows normal distribution. The mean of the distribution $\sqrt{2\chi^2}$ is $\sqrt{2n-1}$ and the standard deviation is equal to 1.

6.4.3 Conditions for applying χ^2 test:

The following conditions should be satisfied before applying χ^2 test.

1. N, the total frequency should be reasonably large, say greater than 50.
2. No theoretical cell-frequency should be less than 5. If it is less than 5, the frequencies should be pooled together in order to make it 5 or more than 5.
3. Each of the observations which makes up the sample for this test must be independent of each other.
4. χ^2 test is wholly dependent on degrees of freedom.

6.5 Testing the Goodness of fit (Binomial and Poisson Distribution):

Karl Pearson in 1900, developed a test for testing the significance of the discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis. This test is known as χ2-test of goodness of fit and is used to test if the deviation between observation (experiment) and theory may be attributed to chance or if it is really due to the inadequacy of the theory to fit the observed data.

Under the null hypothesis that there is no significant difference between the observed and the theoretical values. Karl Pearson proved that the statistic

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(Oi-Ei)^2}{Ei}\right]$$

$$= \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \ldots\ldots\ldots \frac{(O_n-E_n)^2}{E_n}$$

Follows χ^2-distribution with $\nu = n - k - 1$ d.f. where 0_1, 0_2, ...0_n are the observed frequencies, E_1 , E_2..E_n, corresponding to the expected frequencies and k is the number of parameters to be estimated from the given data. A test is done by comparing the computed value with the table value of χ^2 for the desired degrees of freedom.

Example 7:

Four coins are tossed simultaneously and the number of heads occurring at each throw was noted. This was repeated 240 times with the following results.

No. of heads	0	1	2	3	4
No. of throws	13	64	85	58	20

Fit a Binomial distribution assuming under the hypothesis that the coins are unbiased.

Solution:

Null Hypothesis:

H_0**:** The given data fits the Binomial distribution. i.e the coins are unbiased.

$p = q = 1/2$ $n = 4$ $N = 240$

Computation of expected frequencies:

No. of heads	$P(X= x) = 4\, C_x\, p^x\, q^{\,n-x}$	Expected Frequency N. P(X = x)
0	$4C_o \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \left(\frac{1}{16}\right)$	$\left(\frac{1}{16}\right)$ x 240 = 15
1	$4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \left(\frac{4}{16}\right)$	$\left(\frac{4}{16}\right)$ x 240 = 60
2	$4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \left(\frac{6}{16}\right)$	$\left(\frac{6}{16}\right)$ x 240 = 90
3	$4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = \left(\frac{4}{16}\right)$	$\left(\frac{4}{16}\right)$ x 240 = 60
4	$4C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = \left(\frac{1}{16}\right)$	$\left(\frac{1}{16}\right)$ x 240 = 15
		240

Computation of chi square values

Observed Frequency O	Expected Frequency E	$(O-E)^2$	$\left(\frac{(O-E)^2}{E}\right)$
13	15	4	0.27
64	60	16	0.27
85	90	25	0.28
58	60	4	0.07
20	15	25	1.67
			2.56

$$\chi_0^2 = \Sigma\left(\frac{(O-E)^2}{E}\right) = 2.56$$

Expected Value:

$\chi_e^2 = \Sigma\left(\frac{(O-E)^2}{E}\right)$ follows χ^2-distribution with $(n-k-1)$ d.f.

(Here k = 0, since no parameter is estimated from the data)

= 9.488 for $\nu = 5-1 = 4$ d.f.

Inference:

Since $\chi_0^2 < \chi_e^2$ we accept our null hypothesis at 5% level of significance and say that the given data fits Binomial distribution.

Example 8:

The following table shows the distribution of goals in a foot ball match.

No. of goals	0	1	2	3	4	5	6	7
No. of matches	95	158	108	63	40	9	5	2

Fit a Poisson distribution and test the goodness of fit.

Solution:

Null Hypothesis :

The given data fits the Poisson distribution.

Level of significance :

Let $\alpha = 0.05$

Computation of expected frequencies:

$$m = \frac{812}{480} = 1.7$$

$$P(0) = e^{-1.7}\frac{(1.7)^0}{0!} = 0.183.$$

$$f(o) = N.P(o) = 480 \text{ x } 0.183 = 87.84$$

The other expected frequencies will be obtained by using the recurrence formula

$$f(x+1) = \frac{m}{x+1} \text{ x } f(x)$$

Putting x = 0, 1, 2, ... we obtain the following frequencies.

$$f(1) = 1.7 \text{ x } 87.84$$
$$= 149.328$$

$$f(2) = \frac{1.7}{2} \text{ x } 149.328$$
$$= 126.93$$

$$f(3) = \frac{1.7}{3} \text{ x } 126.93$$
$$= 71.927$$

$$f(4) = \frac{1.7}{4} \text{ x } 71.927$$
$$= 30.569$$

$$f(5) = \frac{1.7}{5} \text{ x } 30.569$$
$$= 10.393$$

$$f(6) = \frac{1.7}{6} \text{ x } 10.393$$
$$= 2.94$$

$$f(7) = \frac{1.7}{7} \text{ x } 2.94$$
$$= 0.719$$

No. of goals	0	1	2	3	4	5	6	7	Total
Expected frequency	88	149	127	72	30	10	3	1	480

Computation of statistic:

Observed Frequency O	Expected Frequency E	$(O-E)^2$	$\left(\frac{(O-E)^2}{E}\right)$
95	88	49	0.56
158	150	64	0.43
108	126	324	2.57
63	72	81	1.13
40	30	100	3.33
$\left.\begin{matrix}9\\5\\2\end{matrix}\right\}$ 16	$\left.\begin{matrix}10\\3\\1\end{matrix}\right\}$ 14	4	0.29
			8.31

$$\chi_o^2 = \Sigma\left(\frac{(O-E)^2}{E}\right) = 8.31$$

Expected Value:

$$\chi_e^2 = \Sigma\left(\frac{(O-E)^2}{E}\right) \quad \chi^2 \text{ -distribution with } (n-k-1) \text{ d.f}$$

$$= 9.488 \quad \text{for} \quad \nu = 6-1-1 = 4 \text{ d.f.}$$

Inference:

Since $\chi_0^2 < \chi_e^2$, we accept our null hypothesis at 5% level of significance and say that the given data fits the Poisson distribution.

6.6 Test of independence

Let us suppose that the given population consisting of N items is divided into r mutually disjoint (exclusive) and exhaustive classes $A_1, A_2, \ldots, A_r$ with respect to the attribute A so that randomly selected item belongs to one and only one of the attributes $A_1, A_2, \ldots, A_r$ .Similarly let us suppose that the same population is divided into c mutually disjoint and exhaustive classes $B_1, B_2, \ldots, B_c$ w.r.t another attribute B so that an item selected at random possess one and only one of the attributes $B_1, B_2, \ldots, B_c$. The frequency distribution of the items belonging to the classes A_1, A_2, . .,. A_r and B_1, B_2, . .,. B_c can be represented in the following r × c manifold contingency table.

r × c manifold contingency table

B A	B_1	B_2	…	B_j	…	B_c	Total
A_1	(A_1B_1)	(A_1B_2)	…	(A_1B_j)	…	(A_1B_c)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	…	(A_2B_j)	…	(A_2B_c)	(A_2)
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
A_i	(A_iB_1)	(A_iB_2)	…	(A_iB_j)	…	(A_iB_c)	(A_i)
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
A_r	(A_rB_1)	(A_rB_2)	…	(A_rB_j)	…	(A_rB_c)	(A_r)
Total	(B_1)	(B_2)	…	(B_j)	…	(B_c)	ΣAi = ΣBj = N

(A_i) is the number of persons possessing the attribute A_i, (i=1,2,. .r.), (Bj) is the number of persons possing the attribute B_j,(j=1,2,3,. ..c) and $(A_i\ B_j)$ is the number of persons possessing both the attributes A_i and B_j (i=1,2,. .r. ,j=1,2,. .c).

Also $\Sigma A_i = \Sigma B_j = N$

Under the null hypothesis that the two attributes A and B are independent, the expected frequency for (A_iB_j) is given by

$$= \frac{(Ai)(Bj)}{N}$$

Calculation of statistic:

Thus the under null hypothesis of the independence of attributes,the expected frequencies for each of the cell frequencies of the above table can be obtained on using the formula

$$\chi_0^2 = \Sigma\left(\frac{(O_i - E_i)^2}{E_i}\right)$$

Expected value:

$\chi_e^2 = \Sigma\left(\frac{(O_i - E_i)^2}{E_i}\right)$ follows χ^2-distribution with (r–1) (c–1) d.f

Inference:

Now comparing χ_0^2 with χ_e^2 at certain level of significance ,we reject or accept the null hypothesis accordingly at that level of significance.

6.6.1 2×2 contingency table :

Under the null hypothesis of independence of attributes, the value of χ^2 for the 2×2 contingency table

			Total
	a	b	a+b
	c	d	c+d
Total	a+c	b+d	N

is given by

$$\chi_o^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

6.6.2 Yate’ s correction

In a 2×2 contingency table, the number of d.f. is (2–1)(2–1) =1. If any one of the theoretical cell frequency is less than 5,the use of pooling method will result in d.f = 0 which is meaningless. In this case we apply a correction given by F.Yate (1934) which is usually known as “Yates correction for continuity”. This consisting adding 0.5 to cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly. Thus corrected values of χ^2 is given as

$$\chi^2 = \frac{N\left[(a \text{ « } \frac{1}{2})(d \text{ « } \frac{1}{2}) - (b \pm \frac{1}{2})(c \pm \frac{1}{2})\right]^2}{(a + c)(b + d)(a + b)(c + d)}$$

Example 9:

1000 students at college level were graded according to their I.Q. and the economic conditions of their homes. Use χ^2 test to find out whether there is any association between economic condition at home and I.Q.

Economic Conditions	IQ		Total
	High	Low	
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

Solution:

Null Hypothesis:

There is no association between economic condition at home and I.Q. i.e. they are independent.

$$E_{11} = \frac{(A)(B)}{N} = \frac{600 \times 700}{1000} = 420$$

The table of expected frequencies shall be as follows.

			Total
	420	180	600
	280	120	400
Total	700	300	1000

Observed Frequency O	Expected Frequency E	$(O - E)^2$	$\left(\frac{(O - E)^2}{E}\right)$
460	420	1600	3.81
240	280	1600	5.714
140	180	1600	8.889
160	120	1600	13.333
			31.746

$$\chi_o^2 = \Sigma\left(\frac{(O - E)^2}{E}\right) = 31.746$$

Expected Value:

$\chi_e^2 = \Sigma\left(\frac{(O - E)^2}{E}\right)$ follow χ^2 distribution with (2–1) (2–1) = 1 d.f

$= 3.84$

Inference :

$\chi_o^2 > \chi_e^2$, hence the hypothesis is rejected at 5 % level of significance. $\therefore$ there is association between economic condition at home and I.Q.

Example 10:

Out of a sample of 120 persons in a village, 76 persons were administered a new drug for preventing influenza and out of them, 24 persons were attacked by influenza. Out of those who were not administered the new drug ,12 persons were not affected by influenza.. Prepare

(i) 2x2 table showing actual frequencies.

(ii) Use chi-square test for finding out whether the new drug is effective or not.

Solution:

The above data can be arranged in the following 2 x 2 contingency table.

Table of observed frequencies

New drug	Effect of Influenza		Total
	Attacked	Not attacked	
Administered	24	76 – 24 = 52	76
Not administered	44 –12 = 32	12	120 – 76 = 44
Total	120 – 64 = 56 24 + 32 = 56	52 + 12 = 64	120

Null hypothesis:

‘ Attack of influenza’ and the administration of the new drug are independent.

Computation of statistic:

$$\chi_o^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

$$= \frac{120(24\times12-52\times32)^2}{56\times64\times76\times44}$$

$$= \frac{120(-1376)^2}{56\times64\times76\times44} = \frac{120(1376)^2}{56\times64\times76\times44}$$

= Anti log [log 120 + 2log1376 –(log56 +log64+log76+log44)]

= Antilog (1.2777) = 18.95

Expected value:

$\chi_e^2 = \Sigma\left(\frac{(O - E)^2}{E}\right)$ follows χ^2 distribution with (2–1) × (2–1) d.f

= 3.84

Inference:

Since $\chi_o^2 > \chi_e^2$, H_0 is rejected at 5 % level of significance. Hence we conclude that the new drug is definitely effective in controlling (preventing) the disease (influenza).

Example 11:

Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows

Researchers	No. of Students				Total
	Below average	Average	Above average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are independent?

Solution:

Null Hypothesis:

The sampling techniques adopted by the two researchers are independent.

$$E(86) = \frac{126\times200}{300} = 84$$

$$E(60) = \frac{93\times200}{300} = 62$$

$$E(44) = \frac{69\times200}{300} = 46$$

The table of expected frequencies is given below.

	Below average	Average	Above average	Genius	Total
X	84	62	46	200–192 = 8	200
Y	126 – 84 = 42	93 – 62 = 31	69 – 46 = 23	12 – 8 = 4	100
Total	126	93	69	12	300

Computation of chi-square statistic:

Observed Frequency O	Expected Frequency E	(O – E)	$(O – E)^2$	$\left(\frac{(O - E)^2}{E}\right)$
86	84	2	4	0.048
60	62	–2	4	0.064
44	46	– 2	4	0.087
10	8	2	4	0.500
40	42	–2	4	0.095
33	31	2	4	0.129
$\begin{bmatrix}25\\2\end{bmatrix}27$	$\begin{bmatrix}23\\4\end{bmatrix}27$	0	0	0
300	**300**	**0**		**0.923**

$$\chi_o^2 = \Sigma\left(\frac{(O - E)^2}{E}\right) = 0.923$$

Expected value:

$\chi_e^2 = \Sigma\left(\frac{(O - E)^2}{E}\right)$ follows χ^2distribution with (4–1)(2–1)

= 3 –1 = 2 df

= 5.991

Inference:

Since $\chi_o^2 < \chi_e^2$, we accept the null hypothesis at 5 % level of significance. Hence we conclude that the sampling techniques by the two investigators, do not differ significantly.

6.7 Test for population variance:

Suppose we want to test if the given normal population has a specified variance $\sigma^2 = \sigma_o^2$

Null Hypothesis:

$H_o : \sigma^2 = \sigma_o^2$ if $x_1, x_2 \,..x_n$

Level of significance:

Let α = 5% or 1%

Calculation of statistic:

$$\chi_o^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2}$$

$$\text{Where } s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Expected Value:

$\chi_e^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma_0^2}$ follows χ^2 distribution with (n –1) degrees of freedom.

Inference:

If $\chi_o^2 \leq \chi_e^2$ we accept the null hypothesis otherwise if $\chi_o^2 > \chi_e^2$ we reject the null hypothesis.

Example 12:

A random sample of size 20 from a population gives the sample standard deviation of 6. Test the hypothesis that the population Standard deviation is 9.

Solution:

We are given n = 20 and s = 6

Null hypothesis:

H_0 The population standard deviation is σ= 9.

Level of significance:

Let α = 5 %

Calculation of statistic:

Under null hypothesis H_0:

$$\chi_o^2 = \frac{ns^2}{\sigma^2} = \frac{20\times 36}{9\times 9} = 8.89$$

Expected value:

$$\chi_e^2 = \frac{ns^2}{\sigma^2} \quad \text{follows } \chi^2 \text{ distribution } 20 - 1 = 19 \text{ d.f.}$$

$$= 30.144$$

Inference:

Since $\chi_o^2 < \chi_e^2$, we accept the null hypothesis at 5 % level of significance and conclude that the population standard deviation is 9.

Example 13:

Weights in kgs of 10 students are given below:

38, 40, 45, 53, 47, 43, 55, 48, 52 and 49.

Can we say that the variance of distribution of weights of all the students from which the above sample of 10 students was drawn is equal to 20 sq kg?

Solution:

Null hypothesis :

$H_0 : \sigma^2 = 20$

Computation of sample variance

Weight in Kg x	$x - \overline{x} = x - 47$	$(x - \overline{x})^2$
38	–9	81
40	–7	49
45	–2	4
53	6	36
47	0	0
43	–4	16
55	8	64
48	1	1
52	5	25
49	2	4
		280

Sample mean is

$$\overline{x} = \frac{\sum x}{n} = \frac{470}{10} = 47$$

Calculation of statistic:

Test statistic is

$$\chi_o^2 = \frac{ns^2}{\sigma^2} = \frac{\sum(x - \overline{x})^2}{\sigma^2} = \frac{280}{20} = 14$$

Expected value:

$\chi_e^2 = \frac{ns^2}{\sigma^2}$ follows χ^2 distribution with 10 – 1 = 9 d.f.

$= 16.919$

Inference:

Since$\chi_0^2 < \chi_e^2$ we accept H_o and we conclude that the variance of the distribution of weights of all the students in the population is 20 sq. kgs.

6.8 F – Statistic Definition:

If X is a χ^2 variate with n_1d.f. and Y is an independent χ^2-variate with n_2 d.f., then F - statistic is defined as $F = \frac{X/n_1}{Y/n_2}$

i.e. F - statistic is the ratio of two independent chi-square variates divided by their respective degrees of freedom. This statistic follows G.W. Snedocor' s F-distribution with (n_1, n_2) d.f.

6.8.1 Testing the ratio of variances:

Suppose we are interested to test whether the two normal population have same variance or not. Let x_1, x_2, x_3 .... x_{n_1} , be a random sample of size n_1, from the first population with variance σ_1^2 and y_1, y_2, y_3 ...y_{n_2}, be random sample of size n_2 form the second population with a variance σ_2^2. Obviously the two samples are independent.

Null hypothesis:

$H_0 = \sigma_1^2 = \sigma_2^2 = \sigma^2$

i.e. population variances are same. In other words H_0 is that the two independent estimates of the common population variance do not differ significantly.

Calculation of statistics:

Under H_0, the test statistic is

$$F_0 = \frac{S_1^2}{S_2^2}$$

Where $S_1^2 = \frac{1}{n_1 - 1}\Sigma(x - \bar{x})^2 = \frac{n_1 s_1^2}{n_1 - 1}$

$$S_2^2 = \frac{1}{n_2 - 1}\Sigma(y - \bar{y})^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

It should be noted that numerator is always greater than the denominator in F-ratio

$$F = \frac{\text{Larger Variance}}{\text{Samller Variance}}$$

ν_1 = d.f for sample having larger variance

ν_2 = d.f for sample having smaller variance

Expected value :

$F_e = \frac{S_1^2}{S_2^2}$ follows F- distribution with $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$ d.f

The calculated value of F is compared with the table value for ν_1 and ν_2 at 5% or 1% level of significance If $F_0 > F_e$ then we reject H_0. On the other hand if $F_0 < F_e$ we accept the null hypothesis and it is a inferred that both the samples have come from the population having same variance.

Since F- test is based on the ratio of variances it is also known as the variance Ratio test. The ratio of two variances follows a distribution called the F distribution named after the famous statisticians R.A. Fisher.

Example 14:

Two random samples drawn from two normal populations are :

Sample I:	20	16	26	27	22	23	18	24	19	25		
Sample II:	27	33	42	35	32	34	38	28	41	43	30	37

Obtain the estimates of the variance of the population and test 5% level of significance whether the two populations have the same variance.

Solution:

Null Hypothesis:

H_0: $\sigma_1^2 = \sigma_2^2$ i.e. The two samples are drawn from two populations having the same variance.

Alternative Hypothesis:

H_1: $\sigma_1^2 \neq \sigma_2^2$ (two tailed test)

$$\overline{x_1} = \frac{\sum x_1}{n_1}$$

$$= \frac{220}{10}$$

$$= 22$$

$$\overline{x_2} = \frac{\sum x_2}{n_2}$$

$$= \frac{420}{12}$$

$$= 35$$

x_1	$x_1 - \overline{x_1}$	$(x_1 - \overline{x_1})^2$	x_2	$x_2 - \overline{x_2}$	$(x_2 - \overline{x_2})^2$
20	–2	4	27	–8	64
16	–6	36	33	–2	4
26	4	16	42	7	49
27	5	25	35	0	0
22	0	0	32	–3	9
23	1	1	34	–1	1
18	–4	16	38	3	9
24	2	4	28	–7	49
19	–3	9	41	6	36
25	3	9	43	8	64
220	**0**	**120**	30	–5	25
			37	2	4
			420	**0**	**314**

Level of significance :

0.05

The statistic F is defined by the ratio

$$F_0 = \frac{S_1^2}{S_2^2}$$

Where $S_1^2 = \dfrac{\sum(x_1 - \overline{x_1})^2}{n_1 - 1} = \dfrac{120}{9} = 13.33$

$$S_2^2 = \frac{\sum(x_2 - \overline{x_2})^2}{n_2 - 1} = \frac{314}{11} = 28.54$$

Since $S_2^2 > S_1^2$ larger variance must be put in the numerator and smaller in the denominator

$$\therefore F_0 = \frac{28.54}{13.33} = 2.14$$

Expected value:

$F_e = \dfrac{S_2^2}{S_1^2}$ follows F- distribution with

$\nu_1 = 12–1 = 11$; $\nu_2 = 10–1= 9$ d.f = 3.10

Inference :

Since F_0 < Fe we accept null hypothesis at 5% level of significance and conclude that the two samples may be regarded as drawn from the populations having same variance.

Example 15:

The following data refer to yield of wheat in quintals on plots of equal area in two agricultural blocks A and B Block A was a controlled block treated in the same way as Block B expect the amount of fertilizers used.

	No of plots	Mean yield	Variance
Block A	8	60	50
Block B	6	51	40

Use F test to determine whether variance of the two blocks differ significantly?

Solution:

We are given that

$n_1 = 8$ $n_2 = 6$ $\overline{x}_1 = 60$ $\overline{x}_2 = 51$ $s_1^2 = 50$ $s_2^2 = 40$

Null hypothesis:

$H_0: \sigma_1^2 = \sigma_2^2$ ie there is no difference in the variances of yield of wheat.

Alternative Hypothesis:

$H_1: \sigma_1^2 \neq \sigma_2^2$ (two tailed test)

Level of significance:

Let $\alpha = 0.05$

Calculation of statistic:

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{8 \times 50}{7}$$

$$= 57.14$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{6 \times 40}{5}$$

$$= 48$$

Since $S_1^2 > S_2^2$

$$F_0 = \frac{S_1^2}{S_2^2} = \frac{57.14}{48} = 1.19$$

Expected value:

$F_e = \frac{S_1^2}{S_2^2}$ follows F- distribution with $\nu_1 = 8-1 = 7$ $\nu_2 = 6-1 = 5$ d.f

= 4.88

Inference:

Since $F_0 < F_e$, we accept the null hypothesis and hence infer that there is no difference in the variances of yield of wheat.

Exercise 6

I. Choose the best answer:

1. Student' s ' t' distribution was pioneered by
   (a) Karl Pearson (b) Laplace
   (c)R.A. Fisher (d) William S.Gosset
2. t - distribution ranges from
   (a) $-\infty$ to 0 (b) 0 to ∞ (c) $-\infty$ to ∞ (d) 0 to 1
3. The difference of two means in case of a small samples is tested by the formula
   (a) $t = \frac{\bar{x}_1 - \bar{x}_2}{s}$ (b) $\frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1+n_2}{n_1+n_2}}$
   (c) $t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ (d) $t = \sqrt{\frac{n_1 n_2}{n_1+n_2}}$
4. While testing the significance of the difference between two sample means in case of small samples, the degree of freedom is
   (a) n_1+n_2 (b) n_1+n_2-1
   (c) n_1+n_2-2 (d) n_1+n_2+2
5. Paired t-test is applicable when the observations in the two samples are
   (a) Paired (b) Correlated
   (c) equal in number (d) all the above

6. The mean difference between a paired observations is 15.0 and the standard deviation of differences is 5.0 if n = 9, the value of statistic t is
   (a) 27 (b) 9 (c) 3 (d) zero
7. When observed and expected frequencies completely coincide χ^2 will be
   (a) –1 (b) +1 (c) greater than 1 (d) 0
8. For ν =2, $\chi^2_{0.05}$ equals
   (a) 5.9 (b) 5.99 (c) 5.55 (d) 5.95
9. The calculated value of χ^2 is
   (a) always positive (b) always negative
   (c) can be either positive or negative (d) none of these
10. The Yate' s corrections are generally made when the cell frequency is
   (a) 5 (b) < 5 (c) 1 (d) 4
11. The χ^2 test was derived by
   (a) Fisher (b) Gauss
   (c) Karl Pearson (d) Laplace
12. Degrees of freedom for Chi-square in case of contingency table of order (4 ×3) are
   (a) 12 (b) 9 (c) 8 (d) 6
13. Customarily the larger variance in the variance ratio for F-statistic is taken
   (a) in the denominator (b) in the numerator
   (c) either way (d) none of the above
14. The test statistic F = $\frac{S_1^2}{S_2^2}$ is used for testing
   (a) $H_0: \mu_1 = \mu_2$ (b) $H_0: \sigma_1^2 = \sigma_2^2$
   (c) $H0: \sigma_1 = \sigma_2$ (d) $H_0: \sigma^2 = \sigma_0^2$
15. Standard error of the sample mean in testing the difference between population mean and sample mean under t- statistic
   (a) $\frac{\sigma^2}{\sqrt{n}}$ (b) $\frac{s}{\sqrt{n}}$
   (c) $\frac{\sigma}{\sqrt{n}}$ (d) $\frac{s}{n}$

II. Fill in the blanks:

16. The assumption in t- test is that the population standard deviation is ____________
17. t- values lies in between ____________
18. Paired t- test is applicable only when the observations are ______________
19. Student t- test is applicable in case of ____________ samples
20. The value of χ^2 statistic depends on the difference between ___________ and __________ frequencies
21. The value of χ^2 varies from ____________to ____________
22. Equality of two population variances can be tested by ____________
23. The χ^2 test is one of the simplest and most widely used __________test.
24. The greater the discrepancy between the observed and expected frequency __________the value of χ^2
25. In a contingency table ν __________
26. The distribution of the χ^2 depends on the ___________
27. The variance of the χ^2 distribution is equal to ________ the d.f
28. One condition for application of χ^2 test is that no cell frequency should be __________
29. In a 3×2 contingency table, there are _________ cells
30. F- test is also known as ___________ ratio test.

III. Answer the following

31. Define students ‘ t’ – statistic
32. State the assumption of students ‘ t’ test
33. State the properties of t- distribution
34. What are the applications of t- distribution
35. Explain the test procedure to test the significance of mean in case of small samples.
36. What do you understand by paired ‘ t’ test > What are its assumption.
37. Explain the test procedure of paired – t- test
38. Define Chi square test
39. Define Chi square distribution

40. What is χ^2 test of goodness of fit.

41. What are the precautions are necessary while applying χ^2 test?

42. Write short note on Yate' s correction.

43. Explain the term ' Degrees of freedom'

44. Define non-parametric test

45. Define χ^2 test for population variance

46. Ten flower stems are chosen at random from a population and their heights are found to be (in cms) 63 , 63, 66, 67, 68, 69, 70, 70, 71 and 71. Discuss whether the mean height of the population is 66 cms.

47. A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025cm. A random sample of 10 washers was found to have an average thickness of 0.024cm with a standard deviation of 0.002cm. Test the significance of the deviation.

48. Two types of drugs were used on 5 and 7 patients for reducing their weight.

    Drug A was imported and drug B indigenous. The decrease in the weight after using the drugs for six months was as follows:

Drug A :	10	12	13	11	14		
Drug B :	8	9	12	14	15	10	9

    Is there a significant difference in the efficiency of the two drugs? If not, which drug should you buy?

49. The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records 25 days production. Can you conclude that both the machines are equally efficient at 1% level of significance.

50. A drug is given to 10 patients, and the increments in their blood pressure were recorded to be 3, 6, -2 , +4, –3, 4, 6, 0, 0, 2. Is it reasonable to believe that the drug has no effect on change of blood pressure?

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

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 judges to be a success? Test at 5% level of significance.

52. A survey of 320 families with 5 children each revealed the following distribution.

No of boys	5	4	3	2	1	0
No of Girls	0	1	2	3	4	5
No of Families	14	56	110	88	40	12

Is the result consistent with the hypothesis that the male and female births are equally probable?

53. The following mistakes per page were observed in a book.

No of mistakes per page	0	1	2	3	4	Total
No of pages	211	90	19	5	0	325

Fit a Poisson distribution and test the goodness of fit.

54. Out of 800 persons, 25% were literates and 300 had travelled beyond the limits of their district 40% of the literates were among those who had not travelled. Test of 5% level whether there is any relation between travelling and literacy

55. You are given the following

Fathers	Intelligent Boys	Not intelligent boys	Total
Skilled father	24	12	36
Unskilled Father	32	32	64
Total	56	44	100

Do these figures support the hypothesis that skilled father have intelligent boys?

56. A random sample of size 10 from a normal population gave the following values
    65 , 72, 68, 74, 77, 61,63, 69 , 73, 71
    Test the hypothesis that population variance is 32.
57. A sample of size 15 values shows the s.d to be 6.4. Does this agree with hypothesis that the population s.d is 5, the population being normal.
58. In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7 test whether the difference in the variances is significant at 5% level.
59. The standard deviations calculated from two samples of sizes 9 and 13 are 2.1 and 1.8 respectively. May the samples should be regarded as drawn from normal populations with the same standard deviation?
60. Two random samples were drawn from two normal populations and their values are

A	66	67	75	76	82	84	88	90	92	–	-
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at 5% level of significance.

61. 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 equal to all age groups. The firm takes a random sample from persons attending a pre-view of the new model and obtained the results summarized below:

Person who	Age groups				
	Under 20	20-39	40-50	60 and over	Total
Liked the car	146	78	48	28	300
Disliked the car	54	52	32	62	200
Total	200	130	80	90	500

What conclusions would you draw from the above data?

Answers:

I.

1. (d) 2.(c) 3. (c) 4. (c) 5. (d)
6. (b) 7. (d) 8. (b) 9. (a) 10. (c)
11. (c) 12. (d) 13. (b) 14. (b) 15. (b)

II.

16. not known 17. $-\infty$ to ∞ 18. paired
19 small 20.observed, expected 21. 0, ∞
22. F- test 23.non parametric 24. greater
25. (r–1) ((–1)) 26. degrees of freedom 27. d.f. twice
28. less than 5 29. 6 30. variance

III.

46. t = 1.891 H_0 is accepted 47. t = 1.5 H_0 is accepted
48. t = 0.735 H_0 is accepted 49. t = 7.65 H_0 is rejected
50. t = 2, H_0 is accepted 51. t= 2.58 H_0 is rejected
52. χ^2 = 7.16 H_0 is accepted 53. χ^2 = 0.068 H_0 is accepted
54. χ^2 = 0.0016 H_0 is accepted 55. χ^2 = 2.6 H_0 is accepted
56. χ^2 = 7.3156 H_0 is accepted 57. χ^2 = 24.58 H_0 is rejected
58. χ^2 = 24.576 H_0 is rejected 59. F = 1.41 H_0 is accepted
60. F = 1.415 H_0 is accepted 61. χ^2 = 7.82 , H_0 is rejected

7. ANALYSIS OF VARIANCE

7.0 Introduction:

The analysis of variance is a powerful statistical tool for tests of significance. The term Analysis of Variance was introduced by Prof. R.A. Fisher to deal with problems in agricultural research. The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation where we have three or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population, i.e., they have the same mean. For example, five fertilizers are applied to four plots each of wheat and yield of wheat on each of the plot is given. We may be interested in finding out whether the effect of these fertilizers on the yields is significantly different or in other words whether the samples have come from the same normal population. The answer to this problem is provided by the technique of analysis of variance. Thus basic purpose of the analysis of variance is to test the homogeneity of several means.

Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes and (ii) Chance causes

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be traced separately.

7.1 Definition:

According to R.A. Fisher , Analysis of Variance (ANOVA) is the " Separation of Variance ascribable to one group of causes from the variance ascribable to other group". By this technique the total variation in the sample data is expressed as the sum of its non-negative components where each of these components is a measure of the variation due to some specific independent source or factor or cause.

7.2 Assumptions:

For the validity of the F-test in ANOVA the following assumptions are made.

(i) The observations are independent

(ii) Parent population from which observations are taken is normal and

(iii) Various treatment and environmental effects are additive in nature.

7.3 One way Classification:

Let us suppose that N observations x_{ij} , i = 1, 2, ...k ; j = 1,2...n_i) of a random variable X are grouped on some basis, into k classes of sizes n_1, n_2 , ...n_k respectively ($N = \sum_{i=1}^{k} n_i$) as exhibited below

	Mean	Total
x_{11} x_{12} … x_1n_1	$\overline{x_1}\ .$	$T_{1.}$
x_{21} x_{22} … x_2n_2	$\overline{x_2}\ .$	$T_{2.}$

. . .
. . .
. . .
	.
	.
	.
		.
		.
		.

x_{i1} x_{i2} … x_in_i	$\overline{x_i}\ .$	$T_{i.}$

. . .
. . .
. . .
	.
	.
	.
		.
		.
		.

x_{k1} x_{k2} ..... x_kn_k	$\overline{x_k}\ .$	$T_{k.}$
		G

The total variation in the observation x_{ij} can be spilit into the following two components :

(i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.

(ii) The variation within the classes i.e., the inherent variation of the random variable within the observations of a class.

The first type of variation is due to assignable causes which can be detected and controlled by human endeavour and the second type of variation due to chance causes which are beyond the control of human hand.

In particular, let us consider the effect of *k* different rations on the yield in milk of N cows (of the same breed and stock) divided into k classes of sizes $n_1, n_2, \ldots n_k$ respectively. $N = \sum_{i=1}^{k} n_i$. Hence the sources of variation are

(i) Effect of the rations

(ii) Error due to chance causes produced by numerous causes that they are not detected and identified.

7.4 Test Procedure:

The steps involved in carrying out the analysis are:

1) Null Hypothesis:

The first step is to set up of a null hypothesis

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

Alternative hypothesis H_1: all μ_i ' s are not equal (i = 1,2,..,k)

2) Level of significance : Let α : 0.05

3) Test statistic:

Various sum of squares are obtained as follows.

a) Find the sum of values of all the (N) items of the given data. Let this grand total represented by ' G' .

Then correction factor (C.F) = $\frac{G^2}{N}$

b) Find the sum of squares of all the individual items (x_{ij}) and then the Total sum of squares (TSS) is

$$TSS = \Sigma\Sigma x_{i\,j}^{2} - C.F$$

c) Find the sum of squares of all the class totals (or each treatment total) T_i (i:1,2,..k) and then the sum of squares between the classes or between the treatments (SST) is

$$SST = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C.F$$

Where n_i (i: 1,2,….k) is the number of observations in the i^{th} class or number of observations received by i^{th} treatment

d) Find the sum of squares within the class or sum of squares due to error (SSE) by subtraction.
SSE = TSS - SST

4) Degrees of freedom (d.f):

The degrees of freedom for total sum of squares (TSS) is (N–1). The degrees of freedom for SST is (k–1) and the degrees of freedom for SSE is (N–k)

5) Mean sum of squares:

The mean sum of squares for treatments is $\frac{SST}{k-1}$ and mean sum of squares for error is $\frac{SSE}{N-k}$

6) ANOVA Table

The above sum of squares together with their respective degrees of freedom and mean sum of squares will be summarized in the following table.

ANOVA Table for one-way classification

Sources of variation	d.f	S.S	M.S.S	F ratio
Between treatments	K–1	SST	$\frac{SST}{k-1}$ = MST	$\frac{MST}{MSE} = F_T$
Error	N–k	SSE	$\frac{SSE}{N-k}$ =MSE	
Total	N–1			

Calculation of variance ratio:

Variance ratio of F is the ratio between greater variance and smaller variance, thus

$$F = \frac{\text{Variance between the treatments}}{\text{Variance within the treatment}}$$

$$= \frac{MST}{MSE}$$

If variance within the treatment is more than the variance between the treatments, then numerator and denominator should be interchanged and degrees of freedom adjusted accordingly.

7) Critical value of F or Table value of F:

The Critical value of F or table value of F is obtained from F table for (k-1, N-k) d.f at 5% level of significance.

8) Inference:

If calculated F value is less than table value of F, we may accept our null hypothesis H_0 and say that there is no significant difference between treatments.

If calculated F value is greater than table value of F, we reject our H_0 and say that the difference between treatments is significant.

Example 1:

Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

A	10	12	13	11	10	14	15	13
B	9	11	10	12	13			
C	11	10	15	14	12	13		

Carry out the analysis of variance and state your conclusion.

Solution:

To carry out the analysis of variance, we form the following tables

									Total	Squares
A	10	12	13	11	10	14	15	13	98	9604
B	9	11	10	12	13				55	3025
C	11	10	15	14	12	13			75	5625
								G = 228		

Squares:

A	100	144	169	121	100	196	225	169
B	81	121	100	144	169			
C	121	100	225	196	144	169		
					Total = 2794			

Test Procedure:

Null Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3$

i.e., There is no significant difference between the three processes.

Alternative Hypothesis $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

Level of significance : Let α : 0.05

Test statistic

$$\text{Correct factor (c.f)} = \frac{G^2}{N} = \frac{228^2}{19} = \frac{51984}{19} = 2736$$

$$\text{Total sum of squares (TSS)} = \Sigma\Sigma x_{ij}^2 - \text{C. F} = 2794 - 2736 = 58$$

Sum of squares due to processes = (SST)

$$= \frac{\sum_{i=1}^{3} T_i.^2}{n_i} - \text{C.F}$$

$$= \frac{9604}{8} + \frac{3025}{5} + \frac{5625}{6} - 2736$$

$$= (1200.5 + 605 + 937.5) - 2736$$

$$= 2743 - 2736$$

$$= 7$$

Sum of squares due to error (SSE) = TSS – SST

= 58 – 7 = 51

ANOVA Table

Sources of variation	d.f	S.S	M.S.S	F ratio
Between Processes	3 –1 = 2	7	$\frac{7}{2} = 3.50$	$\frac{3.5}{3.19} = 1.097$
Error	16	51	$\frac{51}{16} = 3.19$	
Total	19 –1 = 18			

Table Value:

Table value of F_e for (2,16) d.f at 5% level of significance is 3.63

Inference:

Since calculated F_0 is less than table value of F_e, we may accept our H_0 and say that there is no significant difference between the three processes.

Example 2:

A test was given to five students taken at random from the fifth class of three schools of a town. The individual scores are

School I	9	7	6	5	8
School II	7	4	5	4	5
School III	6	5	6	7	6

Carry out the analysis of variance

Solution:

To carry out the analysis of variance, we form the following tables.

						Total	Squares
School I	9	7	6	5	8	35	1225
School II	7	4	5	4	5	25	625
School III	6	5	6	7	6	30	900
					Total	G=90	2750

Squares:

School I	81	49	36	25	64
School II	49	16	25	16	25
School III	36	25	36	49	36
				Total = 568	

Test Procedure :

Null Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$ i.e., There is no significant difference between the performance of schools.

Alternative Hypothesis: H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

Level of significance: Let α :0.05

Test Statistic:

Correct factor (c.f) $= \dfrac{G^2}{N}$

$$= \frac{90^2}{15}$$

$$= \frac{8100}{15} = 540$$

Total sum of squares (TSS) = $\Sigma\Sigma x_{ij}^2$ – C. F

$$= 568 - 540 = 28$$

Sum of squares between schools $= \dfrac{\Sigma Ti^2}{n_i} - \text{C.F}$

$$= \frac{2750}{5} - 540$$

$$= 550 - 540 = 10$$

Sum of squares due to error (SSE) = TSS – SST

$$= 28\text{-}10 = 18$$

ANOVA TABLE:

Source of variation	d.f	S.S	M.S.S	F ratio
Between Schools	3-1 = 2	10	$\frac{10}{2} = 5.0$	$\frac{5}{1.5} = 3.33$
Error	12	18	$\frac{18}{12} = 1.5$	
Total	15 -1 = 14			

Table Value:

Table value of F_e for (2,12) d.f at 5% level of significance is 3.8853

Inference:

Since calculated F_0 is less than table value of F_e, we may accept our H_0 and say that there is no significant difference between the performance of schools

7.5 Two way classification:

Let us consider the case when there are two factors which may affect the variate values x_{ij}, e.g the yield of milk may be affected by difference in treatments i.e., rations as well as the difference in variety i.e., breed and stock of the cows. Let us now suppose that the N cows are divided into h different groups or classes according to their breed and stock, each group containing *k* cows and then let us consider the effect of k treatments (i.e., rations given at random to cows in each group) on the yield of milk.

Let the suffix *i* refer to the treatments (rations) and *j* refer to the varieties (breed of the cow), then the yields of milk x_{ij} (i:1,2, ….k; j:1,2..h) of N = h × k cows furnish the data for the comparison of the treatments (rations). The yields may be expressed as variate values in the following k × h two way table.

	Mean	Total
x_{11} x_{12} x_{1j} .. x_{1h}	$\overline{x_1}$.	$T_{1.}$
x_{21} x_{22} x_{2j} .. x_{2h}	$\overline{x_2}$.	$T_{2.}$
. . . .	. .	. .
x_{i1} x_{i2} x_{ij} .. x_{ih}	$\overline{x_i}$.	$T_{i.}$

. . . .
. . . .
. . . .
	.
	.
	.
		.
		.
		.

x_{k1} x_{k2} x_{kj} .. $x_{k}h$	$\overline{x_k}$.	$T_{k.}$
Mean $\overline{x_{\cdot 1}}$. $\overline{x_{\cdot 2}}$ …$\overline{x_{\cdot j}}$ ….$\overline{x_{\cdot h}}$	$\overline{x}$	
Total $T_{\cdot 1}$ $T_{\cdot 2}$. …. $T_{\cdot j}$….$T_{\cdot h}$		G

The total variation in the observation x_{ij} can be split into the following three components:

(i) The variation between the treatments (rations)

(ii) The variation between the varieties (breed and stock)

(iii) The inherent variation within the observations of treatments and within the observations of varieties.

The first two types of variations are due to assignable causes which can be detected and controlled by human endevour and the third type of variation due to chance causes which are beyond the control of human hand.

7.6 Test procedure for Two - way analysis:

The steps involved in carrying out the analysis are:

1. Null hypothesis:

The first step is to setting up a null hypothesis H_0

$H_o : \mu_{1.} = \mu_{2.} = \ldots . \mu_{k.} \;=\; \mu$

$H_o : \mu_{.1} = \mu_{.2} = \ldots \mu_{.h} \;=\; \mu$

i.e., there is no significant difference between rations (treatments) and there is no significant difference between varieties (breed and stock)

2.Level of significance: Let α : 0.05

3.Test Statistic:

Various sums of squares are obtained as follows:

a) Find the sum of values of all the N (k×h) items of the given data. Let this grand total represented by ‘ G’ Then correction factor (C.F) = $\frac{G^2}{N}$

b) Find the sum of squares of all the individual items (x_{ij}) and then the total sum of squares (TSS)
$$\sum_{i-1}^{k} \sum_{j-1}^{k} x^2_{ij} - C.F$$

c) Find the sum of squares of all the treatment (rations) totals, i.e., sum of squares of row totals in the h × k two-way table. Then the sum of squares between treatments or sum of squares between rows is

$$SST = SSR = \sum_{i-1}^{k} \frac{Ti.^2}{h} - C.F$$

where h is the number of observations in each row

d) Find the sum of squares of all the varieties (breed and stock) totals, in the h × k two - way table. Then the sum of squares between varieties or sum of squares between columns is

$SSV = SSC = \frac{\sum_{j-1}^{k} T^2.j}{k} - C.F$ where k is the number of observations in each column.

e) Find the sum of squares due to error by subtraction: i.e., SSE = TSS – SSR – SSC

4. Degrees of freedom:

(i) The degrees of freedom for total sum of squares is N–1 = hk–1

(ii) The degrees of freedom for sum of squares between treatments is k–1

(iii) The degree of freedom for sum of squares between varieties is h – 1

(iv) The degrees of freedom for error sum of squares is (k–1) (h–1)

5. Mean sum of squares (MSS)

(i) Mean sum of squares for treatments (MST) is $\frac{SST}{k-1}$

(ii) Mean sum of squares for varieties (MSV) is $\frac{SSV}{h-1}$

(iii) Mean sum of squares for error (MSE) is $\frac{SSE}{(h-1)(k-1)}$

6. ANOVA TABLE

The above sum of squares together with their respective degrees of freedom and mean sum of squares will be summarized in the following table.

ANOVA Table for Two-way classification

Sources of variation	d.f	SS	MSS	F_0 - ratio
Between Treatments	k-1	SST	MST	$\frac{MST}{MSE} = F_R$
Between Varieties	h-1	SSV	MSV	$\frac{MSV}{MSE} = F_c$
Error	(h-1) (k-1)	SSE	MSE	
Total	N-1			

7. Critical values Fe or Table values of F:

(i) The critical value or table value of ‘ F’ for between treatments is obtained from F table for [(k–1, (k–1) (h–1)] d.f at 5% level of significance.

(ii) The critical value or table value of F_e for between varieties is obtained from F table for [(h–1), (k–1) (h–1)] d.f at 5% level of significance.

8. Inference:

(i) If calculated F_0 value is less than or greater than the table value of F_e for between treatments (rows) H_0 may be accepted or rejected accordingly.

(ii) If calculated F_0 value is less than or greater than the table value of F_e for between varieties (column), H_0 may be accepted or rejected accordingly.

Example 3:

Three varieties of coal were analysed by four chemists and the ash-content in the varieties was found to be as under.

	Chemists			
Varieties	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Carry out the analysis of variance.

Solution:

To carry out the analysis of variance we form the following tables

Chemists						
Varieties	1	2	3	4	Total	Squares
A	8	5	5	7	25	625
B	7	6	4	4	21	441
C	3	6	5	4	18	324
Total	18	17	14	15	**G = 64**	1390
Squares	324	289	196	225	1034	

Individual squares

Chemists				
Varieties	1	2	3	4
A	64	25	25	49
B	49	36	16	16
C	9	36	25	16

Total = 366

Test Procedure:

Null hypothesis:

$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} = \mu$

$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4} = \mu$

(i) i.e., there is no significant difference between varieties (rows)

(ii) i.e., there is no significant difference between chemists (columns)

Alternative hypothesis H_1:

(i) not all $\mu_{i.}$ ' s equal

(ii) not all $\mu_{.j}$' s equal

Level of significance :

Let α : 0.05

Test statistic:

Correction factor (c.f) $= \frac{G^2}{N} = \frac{G^2}{h \times k}$

$$= \frac{(64)^2}{3 \times 4} = \frac{(64)^2}{12}$$

$$= \frac{4096}{12} = 341.33$$

Total sum of squares (TSS) = $\sum_{i-1}^{k} \sum_{j-1}^{k} x^2_{ij} - C.F$

$$= 366 - 341.33$$
$$= 24.67$$

Sum of squares between varieties (Rows)

$$= \frac{\sum T_{i.}{}^2}{4} - C.F$$

$$= \frac{1390}{4} - 341.33$$

$$= 347.5 - 341.33$$
$$= 6.17$$

Sum of squares between chemists (columns)

$$= \frac{\sum T._j{}^2}{3} - C.F$$

$$= \frac{1034}{3} \quad 341.33$$

$$= 344.67 \ - 341.33$$
$$= 3.34$$

Sum of square due to error (SSE)

$$= TSS - SSR - SSC$$
$$= 24.67 - 6.17 - 3.34$$
$$= 24.67 - 9.51$$
$$= 15.16$$

ANOVA TABLE

Sources of variation	d.f	SS	MSS	F ratio
Between Varieties	3 –1 = 2	6.17	3.085	$\frac{3.085}{2.527} = 1.22$
Between Chemists	4 –1 = 3	3.34	1.113	$\frac{2.527}{1.113} = 2.27$
Error	6	15.16	2.527	
Total	12–1 = 11			

Table value :

(i) Table value of F_e for (2,6) d.f at 5% level of significance is 5.14

(ii) Table value of F_e for (6,3) d.f at 5% level of significance is 8.94

Inference:

(i) Since calculated F_0 is less than table value of F_e, we may accept our H_0 for between varieties and say that there is no significant difference between varieties.

(ii) Since calculated F_0 is less than the table value of F_e for chemists, we may accept our H_o and say that there is no significant difference between chemists.

Exercise – 7

I. Choose the best answers:

1. Equality of several normal population means can be tested by
   (a). Bartlet' s test (b) F - test (c) χ^2 -test (d) t- test

2. Analysis of variance technique was developed by
   (a) S. D. Poisson (b) Karl - Pearson
   (c) R.A. Fisher (d) W. S. Gosset

3. Analysis of variance technique originated in the field of
   (a) Agriculture (b) Industry (c) Biology (d) Genetics

4. One of the assumption of analysis of variance is that the population from which the samples are drawn is
   (a) Binomial (b) Poisson (c) Chi-square (d) Normal

5. In the case of one-way classification the total variation can be split into

(a) Two components (b) Three components

(c) Four components (d) Only one component

6. In the case of one-way classification with N observations and t treatments, the error degrees of freedom is

(a) N–1 (b) t –1 (c) N – t (d) Nt

7. In the case of one-way classification with t treatments, the mean sum of squares for treatment is

(a) SST/N–1 (b) SST/ t–1 (c) SST/N–t (d) SST/t

8. In the case of two-way classification with r rows and c columns, the degrees of freedom for error is

(a) (rc) – 1 (b) (r-1).c (c) (r-1) (c-1) (d) (c-1).r

9. In the case of two-way classification, the total variation (TSS) equals.

(a) SSR + SSC + SSE (b) SSR – SSC + SSE

(c) SSR + SSC – SSE (d) SSR + SSC.

10. With 90, 35, 25 as TSS, SSR and SSC respectively in case of two way classification, SSE is

(a) 50 (b) 40 (c) 30 (d) 20

I. Fill in the blanks

11. The technique of analysis of variance was developed by ____________

12. One of the assumptions of Analysis of variance is: observations are ______________

13. Total variation in two – way classification can be split into ____________ components.

14. In the case of one way classification with 30 observations and 5 treatment, the degrees freedom for SSE is _____________

15. In the case of two-way classification with 120, 54, 45 respectively as TSS, SSC, SSE, the SSR is _____________

III. Answer the following:

16. What is analysis of variance?

17. Distinguish between t-test for difference between means and ANOVA.

18. State all the assumptions involved in analysis of variance technique.
19. Explain the structure for one-way classification.
20. Write down the ANOVA table for one-way classification.
21. Distinguish between one - way classification and two-way classification.
22. Explain the structure of two-way classification data.
23. Explain the procedure of obtaining various sums of squares in one-way classification.
24. Write down ANOVA table for two-way classification.
25. Explain the procedure of obtaining various sums of squares in two-way classification.
26. A test was given to a number of students taken at random from the eighth class from each of the 5 schools.

The Individual Scores are:

Schools

I	II	III	IV	V
8	9	12	10	12
9	7	14	11	11
10	11	15	9	10
7	12	12	12	9
8	13	11	10	13

Carry out the analysis of variance and give your conclusions.

27. The following figures relate to production in kg of three varieties A, B and C of wheat shown in 12 plots.

A:	20	18	19		
B:	17	16	19	18	
C:	20	21	20	19	18

Is there any significant difference in the production of the three varieties

28. A special type of fertilizer was used in four agricultural fields A,B,C and D each field was divided into four beds and the fertilizer was applied over them. The respective yields of the beds of four fields are given below. Find whether the difference in mean yields of fields is significant or not?

Plot yield

A	B	C	D
8	9	3	3
12	4	8	7
1	7	2	8
9	1	5	2

29. The following table gives the retail prices of a commodity in (Rs. Per Kg) in some shops selected at random in four cities.

	A	22	24	20	21
CITY	B	20	19	21	22
	C	19	17	21	18
	D	20	22	21	22

Analysis the data to test the significance of the differences between the price of commodity in four cities.

30. For experiments determine the moisture content of sample of a powder, each man taking a sample from each of six consignments Their assessments are:

Consignment

Observer	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Perform an analysis of variance of these data and discuss if there is any significant difference between consignments or between observers.

31. The following are the defective pieces produced by four operators working in turn, on four different machines:

Operator

Machine	I	II	III	IV
A	3	2	3	2
B	3	2	3	4
C	2	3	4	3
D	3	4	3	2

Perform analysis of variance at 5% level of significance to ascertain whether variability in production is due to variability in operator' s performance or variability in machine' s performance.

32. Apply the technique of Analysis of variance to the following data relating to yields of 4 varieties of wheat in 3 blocks.

Blocks

Varieties	1	2	3
I	10	9	8
II	7	7	6
III	8	5	4
IV	5	4	4

33. Four Varieties of potato are planted, each on five plots of ground of the same size and type and each variety is treated with five different fertilizers. The yields in tons are as follows.

Fertilizers

Varieties	F1	F2	F3	F4	F5
V1	1.9	2.2	2.6	1.8	2.1
V2	2.5	1.9	2.2	2.6	2.2
V3	1.7	1.9	2.2	2.0	2.1
V4	2.1	1.8	2.5	2.2	2.5

Perform an analysis of variance and test whether there is any significant difference between yields of different varieties and fertilizers.

34. In an experiment on the effects of temperature conditions in human performance, 8 persons were given a test on 4 temperature conditions. The scores in the test are shown in the following table.

	Persons							
Temperature	1	2	3	4	5	6	7	8
1	70	80	70	90	80	100	90	80
2	70	80	80	90	80	100	90	80
3	75	85	80	95	75	85	95	75
4	65	75	70	85	80	90	80	75

Perform the analysis of variance and state whether there is any significant difference between persons and temperature conditions.

35. The following table gives the number of refrigerators sold by 4 salesmen in three months may, June and July

	Sales Man			
Months	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

Carry out the analysis

Answers

I.

1. b 2. c 3. a 4. d 5. a

6. c 7. b 8. c 9. a 10.c

II.

11. R.A. Fisher 12. Independent 13. Three

14. 25 15. 21

III.

26. Calculated F = 4.56, Table value of F (4,20) = 2.87

27. Calculated F = 9.11, Table value of F (9,2) = 19.39

28. Calculated F = 1.76, Table value of F (12,3) = 8.74

29. Calculated F = 3.29, Table value of F (3,12) = 3.49

30. Calculated F_R = 5.03, Table value of F (3,15) = 3.29

    Calculated F_C = 2.23, Table value of F (5,15) = 2.90

31. Calculated F_R = 2.76, F_C Table value of F (9,3) = 8.81

32. Calculated F_R = 18.23, Table value of F (3,6) = 4.77

    Calculated F_C = 6.4, Table value of F (2,6) = 5.15

33. Calculated F_R = 1.59, Table value of F (3,12) = 3.49

    Calculated F_C = 3.53, Table value of F (4,12) = 3.25

34. Calculated F_R = 3.56, Table value of F (3,21) = 3.07

    Calculated F_C = 14.79, Table value of F (7,21) = 2.49

35. Calculated F_R = 3.33, Table value of F (2,6) = 5.15

    Calculated F_C = 1.02, Table value of F (3,6) = 4.77

8. TIME SERIES

8.0 Introduction:

Arrangement of statistical data in chronological order ie., in accordance with occurrence of time, is known as “Time Series”. Such series have a unique important place in the field of Economic and Business statistics. An economist is interested in estimating the likely population in the coming year so that proper planning can be carried out with regard to food supply, job for the people etc. Similarly, a business man is interested in finding out his likely sales in the near future, so that the businessman could adjust his production accordingly and avoid the possibility of inadequate production to meet the demand. In this connection one usually deal with statistical data, which are collected, observed or recorded at successive intervals of time. Such data are generally referred to as ‘ time series’ .

8.1 Definition:

According to Mooris Hamburg “A time series is a set of statistical observations arranged in chronological order”

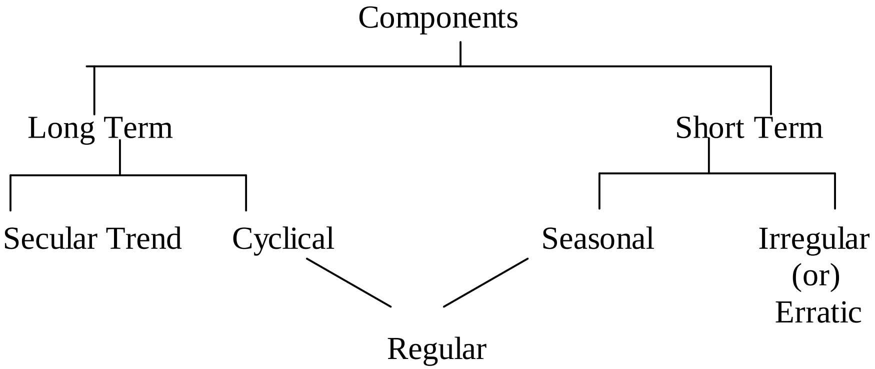
Ya-Lun- chou defining the time series as “A time series may be defined as a collection of readings belonging to different time periods, of some economic variable or composite of variables. A time series is a set of observations of a variable usually at equal intervals of time. Here time may be yearly, monthly, weekly, daily or even hourly usually at equal intervals of time.

Hourly temperature reading, daily sales, monthly production are examples of time series. Number of factors affect the observations of time series continuously, some with equal intervals of time and others are erratic studying, interpreting analyzing the factors is called Analysis of Time Series.

The Primary purpose of the analysis of time series is to discover and measure all types of variations which characterise a time series. The central objective is to decompose the various elements present in a time series and to use them in business decision making.

8.2 Components of Time series:

The components of a time series are the various elements which can be segregated from the observed data. The following are the broad classification of these components.



In time series analysis, it is assumed that there is a multiplicative relationship between these four components. Symbolically,

$$Y = T \times S \times C \times I$$

Where Y denotes the result of the four elements; T = Trend ; S = Seasonal component; C = Cyclical components; I = Irregular component

In the multiplicative model it is assumed that the four components are due to different causes but they are not necessarily independent and they can affect one another.

Another approach is to treat each observation of a time series as the sum of these four components. Symbolically

$$Y = T + S + C + I$$

The additive model assumes that all the components of the time series are independent of one another.

1) Secular Trend or Long - Term movement or simply Trend
2) Seasonal Variation
3) Cyclical Variations
4) Irregular or erratic or random movements(fluctuations)

8.2.1 Secular Trend:

It is a long term movement in Time series. The general tendency of the time series is to increase or decrease or stagnate during a long period of time is called the secular trend or simply trend. Population growth, improved technological progress, changes in consumers taste are the various factors of upward trend. We may notice downward trend relating to deaths, epidemics, due to improved medical facilities and sanitations. Thus a time series shows fluctuations in the upward or downward direction in the long run.

8.2.2 Methods of Measuring Trend:

Trend is measured by the following mathematical methods.

1. Graphical method
2. Method of Semi-averages
3. Method of moving averages
4. Method of Least Squares

Graphical Method:

This is the easiest and simplest method of measuring trend. In this method, given data must be plotted on the graph, taking time on the horizontal axis and values on the vertical axis. Draw a smooth curve which will show the direction of the trend. While fitting a trend line the following important points should be noted to get a perfect trend line.

(i) The curve should be smooth.

(ii) As far as possible there must be equal number of points above and below the trend line.

(iii) The sum of the squares of the vertical deviations from the trend should be as small as possible.

(iv) If there are cycles, equal number of cycles should be above or below the trend line.

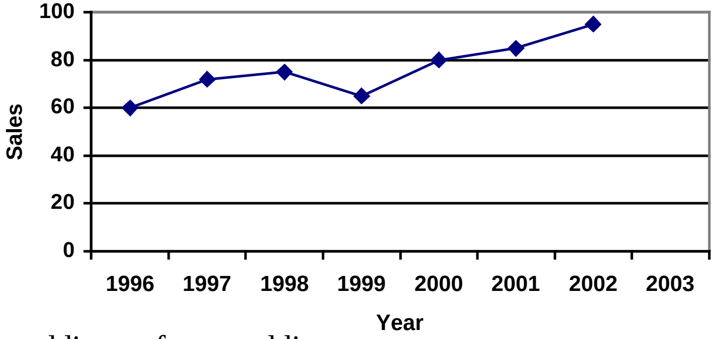
(v) In case of cyclical data, the area of the cycles above and below should be nearly equal.

Example 1:

Fit a trend line to the following data by graphical method.

Year	1996	1997	1998	1999	2000	2001	2002
Sales (in Rs ‘ 000)	60	72	75	65	80	85	95

Solution:



The dotted lines refers trend line

Merits:

1. It is the simplest and easiest method. It saves time and labour.
2. It can be used to describe all kinds of trends.
3. This can be used widely in application.
4. It helps to understand the character of time series and to select appropriate trend.

Demerits:

1. It is highly subjective. Different trend curves will be obtained by different persons for the same set of data.
2. It is dangerous to use freehand trend for forecasting purposes.
3. It does not enable us to measure trend in precise quantitative terms.

Method of semi averages:

In this method, the given data is divided into two parts, preferably with the same number of years. For example, if we are given data from 1981 to 1998 i.e., over a period of 18 years, the two equal parts will be first nine years, i.e., 1981 to 1989 and from 1990 to 1998. In case of odd number of years like 5,7,9,11 etc, two equal parts can be made simply by omitting the middle year. For example, if the data are given for 7 years from 1991 to 1997, the two equal parts would be from 1991 to 1993 and from 1995 to 1997, the middle year 1994 will be omitted.

After the data have been divided into two parts, an average of each parts is obtained. Thus we get two points. Each point is plotted at the mid-point of the class interval covered by respective part and then the two points are joined by a straight line which gives us the required trend line. The line can be extended downwards and upwards to get intermediate values or to predict future values.

Example 2:

Draw a trend line by the method of semi-averages.

Year	1991	1992	1993	1994	1995	1996
Sales Rs in (1000)	60	75	81	110	106	117

Solution:

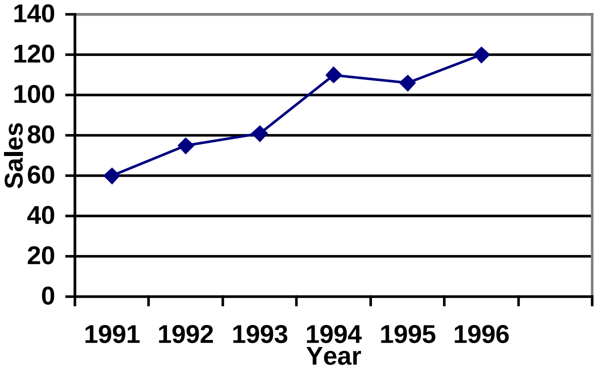
Divide the two parts by taking 3 values in each part.

Year	Sales (Rs)	Semi total	Semi average	Trend values
1991	60			59
1992	75	216	72	72
1993	81			85
1994	110			98
1995	106	333	111	111
1996	117			124

Difference in middle periods = 1995 –1992 = 3 years
Difference in semi averages = 111 –72 = 39
∴ Annual increase in trend = 39/3 = 13

Trend of 1991 = Trend of 1992 -13
= 72-13 = 59
Trend of 1993 = Trend of 1992 +13
= 72 + 13 = 85

Similarly, we can find all the values
The following graph will show clearly the trend line.



Example 3 :

Calculate the trend value to the following data by the method of semi- averages.

Year	1995	1996	1997	1998	1999	2000	2001
Expenditure (Rs in Lakhs)	1.5	1.8	2.0	2.3	2.4	2.6	3.0

Solution:

Year	Expenditure (Rs)	Semi total	Semi average	Trend values
1995	1.5			1.545
1996	1.8	5.3	1.77	1.770
1997	2.0			1.995
1998	2.3			2.220
1999	2.4			2.445
2000	2.6	8.0	2.67	2.670
2001	3.0			2.895

Difference between middle periods = 2000 – 1996
= 4 years
Difference between semi-averages = 2.67 - 1.77
= 0.9

$\therefore$ Annual trend values $= \dfrac{0.9}{4}$

$= 0.225$

Trend of 1995 = Trend of 1996 – 0.225

= 1.77 – 0.225

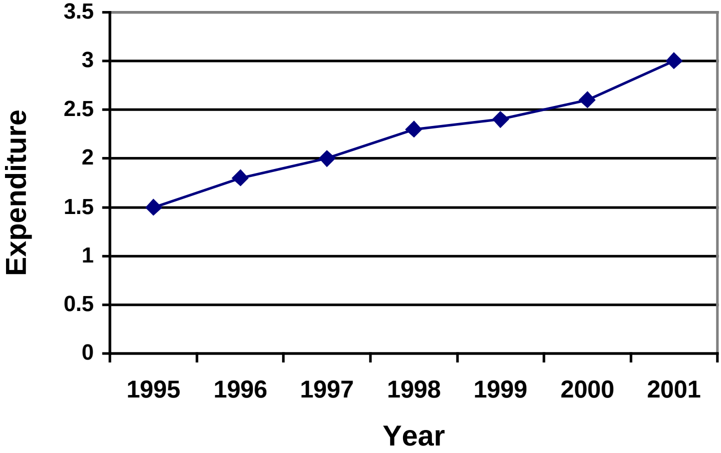
= 1.545

Trend of 1996 = 1.77

Trend of 1997 = 1.77 + 0.225

= 1.995

Similarly we can find all the trend values



Merits:

1. It is simple and easy to calculate
2. By this method every one getting same trend line.
3. Since the line can be extended in both ways, we can find the later and earlier estimates.

Demerits:

1. This method assumes the presence of linear trend to the values of time series which may not exist.
2. The trend values and the predicted values obtained by this method are not very reliable.

Method of Moving Averages:

This method is very simple. It is based on Arithmetic mean. Theses means are calculated from overlapping groups of successive time series data. Each moving average is based on values covering a fixed time interval, called "period of moving average" and is shown against the center of the interval.

The method of ' odd period of moving average is as follows. (3 or 5) . The moving averages for three years is $\frac{a+b+c}{3}$, $\frac{b+c+d}{3}$, $\frac{c+d+e}{3}$ etc

The formula for five yearly moving average is $\frac{a+b+c+d+e}{5}$, $\frac{b+c+d+e+f}{5}$, $\frac{c+d+e+f+g}{5}$ etc.

Steps for calculating odd number of years.

1. Find the value of three years total, place the value against the second year.
2. Leave the first value and add the next three years value (ie 2nd, 3rd and 4th years value) and put it against 3rd year.
3. Continue this process until the last year' s value taken.
4. Each total is divided by three and placed in the next column.

These are the trend values by the method of moving averages

Example 4 :

Calculate the three yearly average of the following data.

Year	1975	1976	1977	1978	1979	1980
Production in (tones)	50	36	43	45	39	38

Year	1981	1982	1983	1984
Production in (tones)	33	42	41	34

Solution:

Year	Production (in tones)	3 years moving total	3 years moving average as Trend values
1975	50	-	-
1976	36	129	43.0
1977	43	124	41.3
1978	45	127	42.3
1979	39	122	40.7
1980	38	110	36.7
1981	33	113	37.7
1982	42	116	38.7
1983	41	117	39.0
1984	34	-	-

Even Period of Moving Averages:

When the moving period is even, the middle period of each set of values lies between the two time points. So we must center the moving averages.

The steps are

1. Find the total for first 4 years and place it against the middle of the 2nd and 3rd year in the third column.
2. Leave the first year value, and find the total of next four-year and place it between the 3rd and 4th year.
3. Continue this process until the last value is taken.
4. Next, compute the total of the first two four year totals and place it against the 3rd year in the fourth column.
5. Leave the first four years total and find the total of the next two four years' totals and place it against the fourth year.
6. This process is continued till the last two four years' total is taken into account.
7. Divide this total by 8 (Since it is the total of 8 years) and put it in the fifth column.

These are the trend values.

Example 5 :

The production of Tea in India is given as follows. Calculate the Four-yearly moving averages

Year	1993	1994	1995	1996	1997	1998
Production (tones)	464	515	518	467	502	540

Year	1999	2000	2001	2002
Production (tones)	557	571	586	612

Solution:

Year	Production (in tones)	4 years Moving total	Total of Two four years	Trend Values
1993	464		-	-
		-		
1994	515			
		1964		
1995	518		3966	495.8
		2002		
1996	467		4029	503.6
		2027		
1997	502		4093	511.6
		2066		
1998	540		4236	529.5
		2170		
1999	557		4424	553.0
		2254		
2000	571		4580	572.5
		2326		
2001	586			
		-		
2002	612			

Merits:

1. The method is simple to understand and easy to adopt as compared to other methods.
2. It is very flexible in the sense that the addition of a few more figures to the data, the entire calculations are not changed. We only get some more trend values.
3. Regular cyclical variations can be completely eliminated by a period of moving average equal to the period of cycles.
4. It is particularly effective if the trend of a series is very irregular.

Demerits:

1. It cannot be used for forecasting or predicting future trend, which is the main objective of trend analysis.
2. The choice of the period of moving average is sometimes subjective.
3. Moving averages are generally affected by extreme values of items.
4. It cannot eliminate irregular variations completely.

8.3 Method of Least Square:

This method is widely used. It plays an important role in finding the trend values of economic and business time series. It helps for forecasting and predicting the future values. The trend line by this method is called the line of best fit.

The equation of the trend line is $y = a + bx$, where the constants a and b are to be estimated so as to minimize the sum of the squares of the difference between the given values of y and the estimate values of y *by* using the equation. The constants can be obtained by solving two normal equations.

$$\Sigma y = na + b\Sigma x \quad .\dots\dots\dots (1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad .\dots\dots (2)$$

Here x represent time point and y are observed values. ‘ n’ is the number of pair- values.

When odd number of years are given

Step 1: Writing given years in column 1 and the corresponding sales or production etc in column 2.

Step 2: Write in column 3 start with 0, 1, 2 .. against column 1 and denote it as X

Step 3: Take the middle value of X as A

Step 4: Find the deviations u = $X - A$ and write in column 4

Step 5: Find u^2 values and write in column 5.

Step 6: Column 6 gives the product uy

Now the normal equations become

$$\Sigma y = na + b\Sigma u \qquad (1) \qquad \text{where } u = X\text{-}A$$

$$\Sigma uy = a\Sigma u + b\Sigma u^2 \qquad (2)$$

Since $\Sigma u = 0$, From equation (1)

$$a = \frac{\Sigma y}{n}$$

From equation (2)

$$\Sigma uy = b\Sigma u^2$$

$$\therefore\ b = \frac{\Sigma uy}{\Sigma u^2}$$

$\therefore$ The fitted straight line is

$$y = a + bu \quad = a + b\,(\,X - A)$$

Example 6:

For the following data, find the trend values by using the method of Least squares

Year	1990	1991	1992	1993	1994
Production (in tones)	50	55	45	52	54

Estimate the production for the year 1996

Solution:

Year (x)	Production (y)	X= x -1990	u = X-A = X-2	u^2	uy	Trend values
1990	50	0	-2	4	-100	50.2
1991	55	1	-1	1	-55	50.7
1992	45	**2 A**	0	0	0	51.2
1993	52	3	1	1	52	51.7
1994	54	4	2	4	108	52.2
Total	256			10	5	

Where A is an assumed value

The equation of straight line is

$$Y = a + bX$$

$$= a + \text{bu}, \text{ where } u = X - 2$$

the normal equations are

$$\Sigma y = na + b\Sigma u \text{.........}(1)$$

$$\Sigma uy = a\Sigma u + \text{b}\Sigma u^2 \text{ ....}(2)$$

since $\Sigma u = 0$ from(1) $\Sigma y = na$

$$a = \frac{\Sigma \text{y}}{\text{n}} = \frac{256}{5} = 51.2$$

From equation (2)

$$\Sigma uy = b\Sigma u^2$$

$$5 = 10\text{b}$$

$$\text{b} = \frac{5}{10} = 0.5$$

The fitted straight line is

$$y = a + bu$$

$$y = 51.2 + 0.5\ (\text{X}-2)$$

$$y = 51.2 + 0.5X - 1.0$$

$$y = 50.2 + 0.5X$$

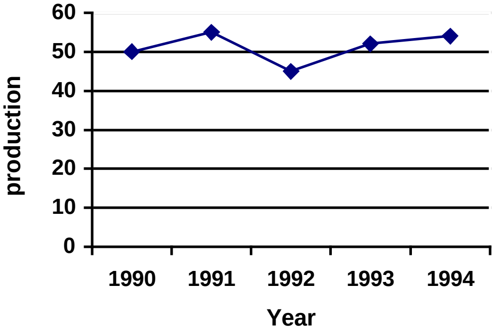
Trend values are, 50.2, 50.7, 51.2, 51.7, 52.2

The estimate production in 1996 is put X = x – 1990

$$\text{X} = 1996 - 1990 = 6$$

$$\text{Y} = 50.2 + 0.5\text{X} = 50.2 + 0.5(6)$$

$$= 50.2 + 3.0 = 53.2 \text{ tonnes.}$$



When **even number of years** are given

Here we take the mean of middle two values of X as A

Then $u = \frac{X - A}{1/2} = 2\ (X\text{-}A)$. The other steps are as given in the odd number of years.

Example 7:

Fit a straight line trend by the method of least squares for the following data.

Year	1983	1984	1985	1986	1987	1988
Sales (Rs. in lakhs)	3	8	7	9	11	14

Also estimate the sales for the year 1991

Solution:

Year (x)	Sales (y)	X = x-1983	u =2X-5	u2	*uy*	Trend values
1983	3	0	-5	25	-15	3.97
1984	8	1	-3	9	-24	5.85
1985	7	2	-1	1	-7	7.73
1986	9	3	1	1	9	9.61
1987	11	4	3	9	33	11.49
1988	14	5	5	25	70	13.37
Total	52		0	70	66	

$$u = \frac{X - A}{1/2}$$

$$= 2\ (X - 2.5) = 2X - 5$$

The straight line equation is

$$y = a + bX = a + bu$$

The normal equations are

$$\Sigma y = na \ \ldots\ldots(1)$$

$$\Sigma uy = b\Sigma u^2 \ \ldots\ldots(2)$$

From (1) $52 = 6a$

$$a = \frac{52}{6}$$

$$= 8.67$$

From (2) 66 = 70 b

$$b = \frac{66}{70}$$

$$= 0.94$$

The fitted straight line equation is

$y = a+bu$

$y = 8.67+0.94(2X-5)$

$y = 8.67 + 1.88X - 4.7$

$y = 3.97 + 1.88X$ -----------(3)

The trend values are

Put $X = 0, y = 3.97$	$X = 1, y = 5.85$
$X = 2, y = 7.73$	$X = 3, y = 9.61$
$X = 4, y = 11.49$	$X = 5, y = 13.37$

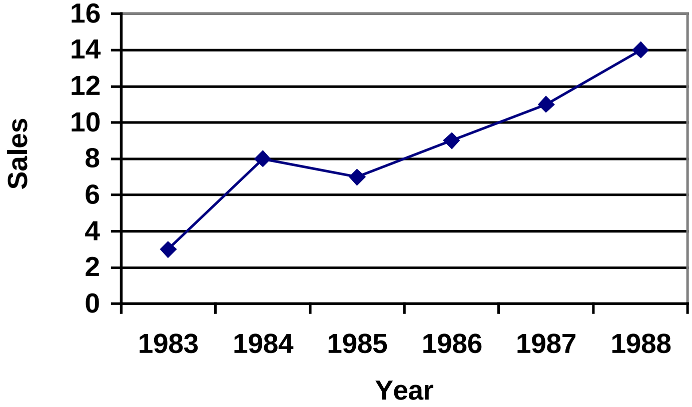
The estimated sale for the year 1991 is; put X = x –1983

$= 1991 – 1983 = 8$

$y = 3.97 + 1.88 \times 8$

$= 19.01$ lakhs

The following graph will show clearly the trend line.



Merits:

1. Since it is a mathematical method, it is not subjective so it eliminates personal bias of the investigator.
2. By this method we can estimate the future values. As well as intermediate values of the time series.
3. By this method we can find all the trend values.

Demerits:

1. It is a difficult method. Addition of new observations makes re-calculations.
2. Assumption of straight line may sometimes be misleading since economics and business time series are not linear.
3. It ignores cyclical, seasonal and irregular fluctuations.
4. The trend can estimate only for immediate future and not for distant future.

8.4 Seasonal Variations:

Seasonal Variations are fluctuations within a year during the season. The factors that cause seasonal variation are

i) Climate and weather condition.

ii) Customs and traditional habits.

For example the sale of ice-creams increase in summer, the umbrella sales increase in rainy season, sales of woolen clothes increase in winter season and agricultural production depends upon the monsoon etc.,

Secondly in marriage season the price of gold will increase, sale of crackers and new clothes increase in festival times.

So seasonal variations are of great importance to businessmen, producers and sellers for planning the future. The main objective of the measurement of seasonal variations is to study their effect and isolate them from the trend.

Measurement of seasonal variation:

The following are some of the methods more popularly used for measuring the seasonal variations.

1. Method of simple averages.
2. Ratio to trend method.
3. Ratio to moving average method.
4. Link relative method

Among the above four methods the method of simple averages is easy to compute seasonal variations.

8.4.1 Method of simple averages

The steps for calculations:

i) Arrange the data season wise

ii) Compute the average for each season.

iii) Calculate the grand average, which is the average of seasonal averages.

iv) Obtain the seasonal indices by expressing each season as percentage of Grand average

The total of these indices would be 100*n* where '*n*' is the number of seasons in the year.

Example 8:

Find the seasonal variations by simple average method for the data given below.

Quarter

Year	I	II	III	IV
1989	30	40	36	34
1990	34	52	50	44
1991	40	58	54	48
1992	54	76	68	62
1993	80	92	86	82

Solution:

Quarter

Year	I	II	III	IV
1989	30	40	36	34
1990	34	52	50	44
1991	40	58	54	48
1992	54	76	68	62
1993	80	92	86	82
Total	238	318	294	270
Average	47.6	63.6	58.8	54
Seasonal Indices	85	113.6	105	96.4

Grand average = $\dfrac{47.6+63.6+58.8+54}{4}$

$$= \frac{224}{4} = 56$$

Seasonal Index for

I quarter $= \dfrac{\textit{First quarterly Average}}{\textit{Grand Average}} \times 100$

$$= \frac{47.6}{56} \times 100 = 85$$

Seasonal Index for

II quarter $= \dfrac{\textit{Second quarterly Average}}{\textit{Grand Average}} \times 100$

$$= \frac{63.6}{56} \times 100 = 113.6$$

Seasonal Index for

III quarter $= \dfrac{\textit{Third quarterly Average}}{\textit{Grand Average}} \times 100$

$$= \frac{58.8}{56} \times 100 = 105$$

Seasonal Index for

IV quarter $= \dfrac{\textit{Fourth quarterly Average}}{\textit{Grand Average}} \times 100$

$$= \frac{54}{56} \times 100 = 96.4$$

Example 9:

Calculate the seasonal indices from the following data using simple average method.

Year

Quarter	1974	1975	1976	1977	1978
I	72	76	74	76	74
II	68	70	66	74	74
III	80	82	84	84	86
IV	70	74	80	78	82

Solution:

Quarter

Year	I	II	III	IV
1974	72	68	80	70
1975	76	70	82	74
1976	74	66	84	80
1977	76	74	84	78
1978	74	74	86	82
Total	372	352	416	384
Average	74.4	70.4	83.2	76.8
Seasonal Indices	97.6	92.4	109.2	100.8

$$\text{Grand Average} = \frac{74.4+70.4+83.2+76.8}{4}$$

$$= \frac{304.8}{4} = 76.2$$

Seasonal Index for

$$\text{I quarter} = \frac{\textit{First quarterly Average}}{\textit{Grand Average}} \times 100$$

$$= \frac{74.4}{76.2} \times 100$$

$$= 97.6$$

Seasonal Index for

$$\text{II quarter} = \frac{\textit{Second quarterly Average}}{\textit{Grand Average}} \times 100$$

$$= \frac{70.4}{76.2} \times 100$$

$$= 92.4$$

Seasonal Index for

$$\text{III quarter} = \frac{\textit{Third quarterly Average}}{\textit{Grand Average}} \times 100$$

$$= \frac{83.2}{76.2} \times 100$$

$$= 109.2$$

Seasonal Index for

IV quarter $= \dfrac{\textit{Fourth quarterly Average}}{\textit{Grand Average}} \times 100$

$$= \frac{76.8}{76.2} \times 100$$

$$= 100.8$$

The total of seasonal indices calculated must be equal to 400 here we have = 97.6 + 92.4 + 109.2 + 100.8

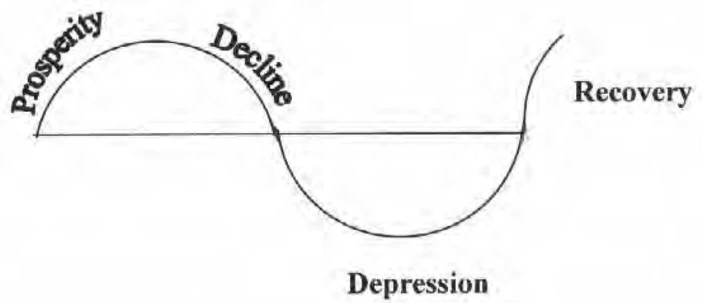
= 400 hence verified.

Cyclical variations:

The term cycle refers to the recurrent variations in time series, that extend over longer period of time, usually two or more years. Most of the time series relating to economic and business show some kind of cyclic variation. A business cycle consists of the recurrence of the up and down movement of business activity. It is a four-phase cycle namely.

1. Prosperity 2. Decline 3. Depression 4. Recovery

Each phase changes gradually into the following phase. The following diagram illustrates a business cycle.



The study of cyclical variation is extremely useful in framing suitable policies for stabilising the level of business activities. Businessmen can take timely steps in maintaining business during booms and depression.

Irregular variation:

Irregular variations are also called erratic. These variations are not regular and which do not repeat in a definite pattern.

These variations are caused by war, earthquakes, strikes flood, revolution etc. This variation is short-term one, but it affect all the components of series. There is no statistical techniques for measuring or isolating erratic fluctuation. Therefore the residual that remains after eliminating systematic components is taken as representing irregular variations.

FORECASTING

8.5 Introduction:

A very important use of time series data is towards forecasting the likely value of variable in future. In most cases it is the projection of trend fitted into the values regarding a variable over a sufficiently long period by any of the methods discussed latter. Adjustments for seasonal and cyclical character introduce further improvement in the forecasts based on the simple projection of the trend. The importance of forecasting in business and economic fields lies on account of its role in planning and evaluation. If suitably interpreted, after consideration of other forces, say political, social governmental policies etc., this statistical technique can be of immense help in decision making.

The success of any business depends on its future estimates. On the basis of these estimates a business man plans his production stocks, selling market, arrangement of additional funds etc. Forecasting is different from predictions and projections. Regression analysis, time series analysis, Index numbers are some of the techniques through which the predictions and projections are made. Where as forecasting is a method of foretelling the course of business activity based on the analysis of past and present data mixed with the consideration of ensuring economic policies and circumstances. In particularly forecasting means fore-warning. Forecasts based on statistical analysis are much reliable than a guess work.

According to T.S.Levis and and R.A. Fox, “ Forecasting is using the knowledge we have at one time to estimate what will happen at some future movement of time”.

8.5.1 Methods of Business forecasting:

There are three methods of forecasting

1. Naive method
2. Barometric methods
3. Analytical Methods

1. Naive method :

It contains only the economic rhythm theory.

2. Barometric methods:

It covers

i) Specific historical analogy
ii) Lead- Lag relationship
iii) Diffusion method
iv) Action –reaction theory

3. Analytical Methods:

It contains

i) The factor listing method
ii) Cross-cut analysis theory
iii) Exponential smoothing
iv) Econometric methods

The economic rhythm theory:

In this method the manufactures analysis the time-series data of his own firm and forecasts on the basis of projections so obtained. This method is applicable only for the individual firm for which the data are analysed, The forecasts under this method are not very reliable as no subjective matters are being considered.

Diffusion method of Business forecasting

The diffusion index method is based on the principle that different factors, affecting business, do not attain their peaks and troughs simultaneously. There is always time-log between them. This method has the convenience that one has not to identify which series has a lead and which has a lag. The diffusion index depicts the movement of broad group of series as a whole without bothering about the individual series. The diffusion index shows the percentage of a given set of series as expanding in a time period. It should be carefully noted that the peaks and troughs of diffusion index are not the peaks troughs of the business cycles. All series do not expand or contract concurrently. Hence if more than 50% are expanding at a given time, it is taken that the business is in the process of booming and vice - versa.

The graphic method is usually employed to work out the diffusion index. The diffusion index can be constructed for a group of business variables like prices, investments, profits etc.

Cross cut analysis theory of Business forecasting:

In this method a thorough analysis of all the factors under present situations has to be done and an estimate of the composite effect of all the factors is being made. This method takes into account the views of managerial staff, economists, consumers etc. prior to the forecasting. The forecasts about the future state of the business is made on the basis of over all assessment of the effect of all the factors.

Exercise – 8

I. Choose the best answer:

1. A time series consists of
   a) Two components b) Three Components
   c) Four components d) Five Components
2. Salient features responsible for the seasonal variation are
   a) Weather b) Social customers
   c) Festivals d) All the above
3. Simple average method is used to calculate
   a) Trend Values b) Cyclic Variations
   c) Seasonal indices d) None of these
4. Irregular variations are
   a) Regular b) Cyclic
   c) Episodic d) None of the above
5. If the slope of the trend line is positive it shows
   a) Rising Trend b) Declining trend
   c) Stagnation d) None of the above
6. The sales of a departmental store on Diwali are associated with the component of time-series
   a) Secular trend b) Seasonal variation
   c) Irregular variation d) All the above

7. The component of time-series attached to long term variation is termed as
   a) Secular Trend b) Seasonal Variation
   c) Irregular variation d) Cyclic variation
8. Business forecasts are made on the basis of
   a) Present Data b) Past data
   c) Polices and circumstances d) All the above
9. Econometric methods involve
   a) Economics and mathematics b) Economics and Statistics
   c) Economics, Statistics and Mathematics
   d) None of the above
10. The economic rhythm theory comes under the category of
   a) Analytical methods b) Naive method
   c) Barometric methods d) None of the above

II. Fill in the blanks:

11. A time series in a set of values arranged in ________ order
12. Quarterly fluctuations observed in a time series represent _______ variation
13. Periodic changes in a business time series are called _________
14. A complete cycle passes through _____________stages of phenomenon.
15. An overall tendency of rise and fall in a time series represents ___________
16. The trend line obtained by the method of least square is known as the ___________
17. Forecasting is different from ________ and _________.
18. No statistical techniques measuring or isolating _________

III. Answer the following questions

19. What is a time series?
20. What are the components of time series.
21. Write briefly about seasonal variation.
22. What is cyclic variation.
23. Give the names of different methods of measuring trend.
24. What are the merits and demerits of the semi-average method.
25. Discuss the mathematical models for a time series analysis.

26. Discuss irregular variation in the context of time series.
27. What do you understand by business fore-casting
28. Give the names of different methods of fore casting.
29. Write briefly about any one method of forecasting?
30. In what sense forecasting differ from prediction and projection.

IV. Problems

31. With the help of graph paper obtain the trend values.

Year	1996	1997	1998	1999	2000	2001	2002
Value	65	85	95	75	100	80	130

32. Using graphical method, fit a trend-line to the following data.

Year	1982	1983	1984	1985	1986	1987
Value	24	22	25	26	27	26

33. Draw a trend line by the method of semi-averages.

Year	1993	94	95	96	97	98	99	2000
Sales	210	200	215	205	220	235	210	235

34. The following figures are given relating to the output in a factory. Draw a trend-line with the help of method of semi-averages.

Year	1996	1997	1998	1999	2000	2001	2002
Output	600	800	1000	800	1200	1000	1400

35. Calculate three yearly moving average of the following data

Year	91	92	93	94	95	96	97	98	99	00
No of students	15	18	17	20	23	25	29	33	36	40

36. The following figures relating to the profits of a commercial concern for 8 years. Find the 3-yearly moving averages.

Years	Profits	Years	Profits
1995	15,420	1999	26,120
1996	14,470	2000	31,950
1997	15,520	2001	35,370
1998	21,020	2002	35,670

37. Construct a four yearly centered moving average from the following data.

Year	1940	1950	1960	1970	1980	1990	2000
Imported cotton consumption (‘ 000)	129	131	106	91	95	84	93

38. From the following data calculate the 4-yearly moving average and determine the trend values. Find the short-term fluctuations. Plot the original data and the trend on a graph.

Year	93	94	95	96	97	98	99	00	01	02
Value	50	36.5	43	44.5	38.9	38.1	32.6	41.7	41.1	33.8

39. Calculate trend value by taking 5 yearly period of moving average from the data given below

Year	1987	88	89	90	91	92	93	94
Production in tones	4	5	6	7	9	6	5	7

Year	95	96	97	98	99	2000	01	02
Production in tones	8	7	6	8	9	10	7	9

40. Fit a straight line trend by the method of least squares to the following data and calculate trend values.

Year	1996	1997	1998	1999	2000
Sales of TV Sets (Rs ‘ 000)	4	6	7	8	10

Estimate the sales for the year 2005

41. Below are given the figures of production in ‘ 000 quintals of a sugar factory.

Year	1994	95	96	97	98	99	2000
Production in tones	80	90	92	83	94	99	92

42. Fit a straight line trend by the method of least square to the following data.

Year	1996	97	98	99	2000	2001
Profit	300	700	600	800	900	700

43. Fit a straight line trend by the method of least squares to the following data. Estimate the earnings for the year 2002.

Year	1993	94	95	96	97	98	99	2000
Earnings	38	40	65	72	69	60	87	95

44. Compute the average seasonal movement for the following series.

Year	Ist quarter	IInd quarter	IIIrd quarter	IVth quarter
1999	3.5	3.9	3.4	3.6
2000	3.5	4.1	3.7	4.0
2001	3.5	3.9	3.7	4.2
2002	4.0	4.6	3.8	4.5
2003	4.1	4.4	4.2	4.5

45. Obtain seasonal fluctuations from the following time-series Quarterly output of coal for four years.

Year	2000	2001	2002	2003
I	65	58	70	60
II	58	63	59	55
III	56	63	56	51
IV	61	67	52	58

Answers

I.

1. (c) 2.(d) 3. (c) 4. (c) 5. (a) 6. (b)
7. (a) 8.(d) 9. (c) 10.(b)

II.

11. Chronological 12. Seasonal 13. Cycles
14. four 15. secular trend
16. line of best fit 17. Prediction, projection
18. Erratic Fluctuation.

IV.

33. Trend values are 200.94, 205.31, 209.69, 214.06, 218.43,
222.80, 227.19, 231.56

34. 700, 800, 900, 1000, 1100, 1200, 1300

35. 16.7, 18.3, 20, 22.7, 25.7, 29, 32.7, 36.3

36. 15137, 17003, 20363, 26363, 31.147, 34330

37. 110.0, 99.88, 92.38

38. 42.1, 40.9, 39.8, 38.2, 38.1, 37.8,

39. 6.2, 6.6, 6.6, 6.8, 7.0, 6.6, 6.6, 7.2, 7.6, 8.0, 8.0

40. Trend values are 4.2, 5.6, 7; 8.4, 9.8

41. 84, 86, 88, 90, 92, 94, 96

42. 446.67, 546.67, 626.67, 706.67, 786.67, 866.67

43. 40.06, 47.40, 54.74, 62.08, 69.42, 76.76, 84.10, 91.44

44. 94.18, 105.82, 95.19, 105.32

45. 106.4, 98.7, 94.9,100

9. THEORY OF ATTRIBUTES

9.0 Introduction:

Generally statistics deal with quantitative data only. But in behavioural sciences, one often deals with the variable which are not quantitatively measurable. Literally an attribute means a quality on characteristic which are not related to quantitative measurements. Examples of attributes are health, honesty, blindness etc. They cannot be measured directly. The observer may find the presence or absence of these attributes. Statistics of attributes based on descriptive character.

9.1 Notations:

Association of attribute is studied by the presence or absence of a particular attribute. If only one attribute is studied, the population is divided into two classes according to its presence or absence and such classification is termed as division by dichotomy. If a class is divided into more than two scale-classes, such classification is called manifold classification.

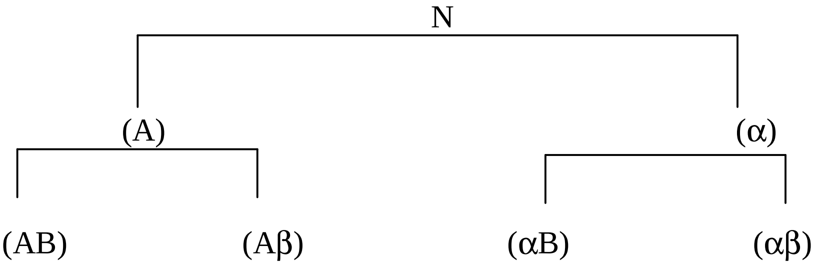
Positive class which denotes the presence of attribute is generally denoted by Roman letters generally A,B,…etc and the negative class denoting the absence of the attribute and it is denoted by the Greek letters α, β…etc For example, A represents the attribute 'Literacy' and B represents 'Criminal'. α and β represents the 'Illiteracy' and 'Not Criminal' respectively.

9.2 Classes and Class frequencies:

Different attributes, their sub-groups and combinations are called different classes and the number of observations assigned to them are called their class frequencies.

If two attributes are studied the number of classes will be 9. (i.e.,) (A) , (α), (B), (β), (A β) (α β), (α B) and N.

The chart given below illustrate it clearly.



The number of observations or units belonging to class is known as its frequency are denoted within bracket. Thus (A) stands for the frequency of A and (AB) stands for the number objects possessing the attribute both A and B. The contingency table of order (2×2) for two attributes A and B can be displayed as given below

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	N

Relationship between the class frequencies:

The frequency of a lower order class can always be expressed in terms of the higher order class frequencies.

i.e., $N = (A) + (\alpha) = (B) + (\beta)$

$(A) = (AB) + (A\beta)$

$(\alpha) = (\alpha B) + (\alpha\beta)$

$(B) = (AB) + (\alpha B)$

$(\beta) = (A\beta) + (\alpha \beta)$

If the number of attributes is n, then there will be 3^n classes and we have 2^n cell frequencies.

9.3 Consistency of the data:

In order to find out whether the given data are consistent or not we have to apply a very simple test. The test is to find out whether any or more of the ultimate class-frequencies is negative or not. If none of the class frequencies is negative we can safely calculate that the given data are consistent (i.e the frequencies do not conflict in any way each other). On the other hand, if any of the ultimate class frequencies comes to be negative the given data are inconsistent.

Example 1:

Given N = 2500, (A) = 420, (AB) = 85 and (B) = 670. Find the missing values.

Solution:

We know $N = (A) + (\alpha) = (B) + (\beta)$

$(A) = (AB) + (A\beta)$

$(\alpha) = (\alpha B) + (\alpha\beta)$

$(B) = (AB) + (\alpha B)$

$(\beta) = (A\beta) + (\alpha \beta)$

From (2) $420 = 85 + (A\beta)$

$\therefore \quad (A\beta) = 420 - 85$

$(A\beta) = 335$

From (4) $670 = 85 + (\alpha B)$

$\therefore \quad (\alpha B) = 670 - 85$

$(\alpha B) = 585$

From (1) $2500 = 420 + (\alpha)$

$\therefore \quad (\alpha) = 2500 - 420$

$(\alpha) = 2080$

From (1) $(\beta) = 2500 - 670$

$(\beta) = 1830$

From (3) $= 2080 = 585 + (\alpha\beta)$

$\therefore (\alpha\beta) = 1495$

Example 2:

Test the consistency of the following data with the symbols having their usual meaning.

N = 1000 (A) = 600 (B) = 500 (AB) = 50

Solution:

	A	α	Total
B	50	450	500
β	550	-50	500
Total	600	400	1000

Since (αβ)) = –50, the given data is inconsistent.

Example 3:

Examine the consistency of the given data. N = 60 (A) = 51 (B) = 32 (AB) = 25

Solution:

	A	α	Total
B	25	7	32
β	26	2	28
Total	51	9	60

Since all the frequencies are positive, it can be concluded that the given data are consistent.

9.4 Independence of Attributes:

If the attributes are said to be independent the presence or absence of one attribute does not affect the presence or absence of the other. For example, the attributes skin colour and intelligence of persons are independent.

If two attributes A and B are independent then the actual frequency is equal to the expected frequency

$$(AB) = \frac{(A).(B)}{N}$$

Similarly $$(\alpha\ \beta) = \frac{(\alpha).(\beta)}{N}$$

9.4.1 Association of attributes:

Two attributes A and B are said to be associated if they are not independent but they are related with each other in some way or other.

The attributes A and B are said to be positively associated if

$$(AB) > \frac{(A).(B)}{N}$$

If $(AB) < \dfrac{(A).(B)}{N}$,. then they are said to be negatively associated.

Example 4:

Show that whether A and B are independent, positively associated or negatively associated.

$(AB) = 128$, $(\alpha B) = 384$, $(A\beta) = 24$ and $(\alpha\beta) = 72$

Solution:

$$
\begin{aligned}
(A) &= (AB) + (A\beta) \\
&= 128 + 24 \\
(A) &= 152 \\
(B) &= (AB) + (\alpha B) \\
&= 128 + 384 \\
(B) &= 512 \\
(\alpha) &= (\alpha B) + (\alpha\beta) \\
&= 384 + 72 \\
\therefore (\alpha) &= 456 \\
(N) &= (A) + (\alpha) \\
&= 152 + 456 \\
&= 608
\end{aligned}
$$

$$\frac{(A)\times(B)}{N} = \frac{152\times512}{608}$$

$$= 128$$

$$(AB) = 128$$

$$\therefore (AB) = \frac{(A)\times(B)}{N}$$

Hence A and B are independent

Example 5:

From the following data, find out the types of association of A and B.

1) N = 200	(A) = 30	(B) = 100	(AB) = 15
2) N = 400	(A) = 50	(B) = 160	(AB) = 20
3) N = 800	(A) = 160	(B) = 300	(AB) = 50

Solution:

1. Expected frequency of (AB) $= \dfrac{(A).(B)}{N}$

$$= \frac{(30)(100)}{200} = 15$$

Since the actual frequency is equal to the expected frequency, ie 15 = 15, therefore A and B are independent.

2. Expected frequency of (AB) $= \dfrac{(A).(B)}{N}$

$$= \frac{(50)(160)}{400} = 20$$

Since the actual frequency is greater than expected frequency. i.e., 25 > 20, therefore A and B are positively associated.

3. Expected frequency of (AB) $= \dfrac{(A).(B)}{N} = \dfrac{(160)(300)}{800} = 60$

Since Actual frequency is less than expected frequency i.e., 50 < 60 therefore A and B are negatively associated.

9.5 Yules' co-efficient of association:

The above example gives a rough idea about association but not the degree of association. For this Prof. G. Undy Yule has suggested a formula to measure the degree of association. It is a relative measure of association between two attributes A and B.

If (AB), (αB), (Aβ) and (αβ) are the four distinct combination of A, B, α and β then Yules' co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)}$$

Note:

I. If Q = +1 there is perfect positive association

If Q = -1 there is perfect negative association

If Q = 0 there is no association (ie) A and B are independent

1. For rememberance of the above formula , we use the table below

	A	α
B	AB	αB
β	Aβ	αβ

Example 6:

Investigate the association between darkness of eye colour in father and son from the following data.

Fathers' with dark eyes and sons' with dark eyes	= 50
Fathers' with dark eyes an sons' with no dark eyes	= 79
Fathers' with no dark eyes and sons with dark eyes	= 89
Neither son nor father having dark eyes	= 782

Solution:

Let A denote the dark eye colour of father and B denote dark eye colour of son.

	A	α	Total
B	50	89	139
β	79	782	861
Total	129	871	1000

Yules' co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta)-(A\beta).(\alpha B)}{(AB)(\alpha\beta)+(A\beta).(\alpha B)}$$

$$= \frac{50\times782-79\times89}{50\times782+79\times89}$$

$$= \frac{32069}{46131} = 0.69$$

$\therefore$ there is a positive association between the eye colour of fathers' and sons' .

Example 7 :

Can vaccination be regarded as a preventive measure of small pox from the data given below.

Of 1482 persons in a locality, exposed to small pox, 368 in all were attacked, among the 1482 persons 343 had been vaccinated among these only 35 were attacked.

Solution:

Let A denote the attribute of vaccination and B denote that of attacked.

	A	α	Total
B	35	333	368
β	308	806	1114
Total	343	1139	1482

Yules' co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta)-(A\beta).(\alpha B)}{(AB)(\alpha\beta)+(A\beta).(\alpha B)}$$

$$= \frac{35\times806-308\times333}{35\times806+308\times333}$$

$$= \frac{-74354}{130774} = -0.57$$

i.e., there is a negative association between attacked and vaccinated. In other words there is a positive association between not attacked and vaccinated. Hence vaccination can be regarded as a preventive measure for small pox.

Example 8:

In a co-educational institution, out of 200 students, 150 were boys. They took an examination and it was found that 120 passed, 10 girls failed. Is there any association between sex and success in the examination.

Solution:

Let A denote boys and α denote girls. Let B denote those who passed the examination and β denote those who failed.

We have given N = 200 (A) = 150 (AB) = 120 (αβ) = 10

Other frequencies can be obtained from the following table

	A	α	Total
B	120	40	160
β	30	10	40
Total	150	50	200

Yule' s co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)}$$

$$= \frac{120 \times 10 - 30 \times 40}{120 \times 10 + 30 \times 40} = 0$$

Therefore, there is no association between sex and success in the examination.

Recall

(A) (B) denote positive attributes

(α) (β) denote negative attributes

2 ×2 contingency table.

X	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	N

Vertical Total	Horizontal Total
$(AB) + (A\beta) = (A)$	$(AB) + (\alpha B) = B$
$(\alpha B) + (\alpha\beta) = (\alpha)$	$(A\beta) + (\alpha\beta) = \beta$
$(A) + (\alpha) = N$	$(B) + (\beta) = N$

Types of Association

Positive Association if $(AB) > \frac{(A).(B)}{N}$

Negative Association if $(AB) < \frac{(A).(B)}{N}$

Independent if $(AB) = \frac{(A).(B)}{N}$

Yule' s co-efficient of Association

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)}$$

Exercise – 9

I. Choose the best answer:

1. Measures of association in usually deal with
   (a) Attributes (b) Quantitative factors
   (c) Variables (d) Numbers
2. The frequency of class can always be expressed as a sum of frequencies of
   (a) Lower order classes (b) Higher order classes
   (c) Zero order classes (d) None of the above
3. With the two attributes the total number of class frequencies is
   (a) Two (b) Four (c) Eight (d) Nine
4. If for two the attributes are A and B, $(AB) > \frac{(A).(B)}{N}$ the attributes are
   (a) Independent (b) Positively associated
   (c) Negatively associated (d) No conclusion

5. In case of two attributes A and B the class frequency (AB) = 0 the value of Q is
   (a) 1 (b) – 1 (c) 0 (d) $-1 \leq Q \leq 1$

II. Fill in the blanks:

6. If an attribute has two classes it is said to be ____________
7. In case of consistent data, no class frequency can be _________
8. If A and B are independent Yule' s co-efficient is equal to ________
9. If A and B are negatively associated then __________
10. If N = 500, (A) = 300, (B) = 250 and (AB) = 40 the data are ________

III. Answer the following:

11. Give a brief idea of notations used in classification of attributes
12. How can the frequencies for various attributes be displayed in contingency table
13. What do you understand by consistency of data.
14. Write briefly about association of attributes.
15. Give Yule' s co-efficient of association

IV. Problems

16. For two attributes A and B, we have (AB) = 35, (A) = 55; N=100 and (B) = 65. Calculate the missing values.
17. From the following ultimate class frequencies, find the frequencies of positive and negative classes and the total number of observations. (AB) = 9, $(A\beta)$ = 14, (αB) = 4 and $(\alpha\beta)$ = 37
18. Verify whether the given data N = 100, (A) = 75, (B) = 60 and (AB) = 15 are consistent.
19. Find whether A and B are independent in the following data
    (AB) = 256 (αB) = 768 $(A\beta)$ = 48 $(\alpha\beta)$ = 144
20. In a report on consumer' s preference it was given that out of 500 persons surveyed 410 preferred variety A 380 preferred

variety B and 270 persons linked both. Are the data consistent?

21. For two attributes A and B, we have (AB) = 35, (A) = 55, N=100, ($\alpha\beta$) = 20. Calculate the Yule' s co-efficient of association.

22. Given N = 1500, (A) = 383, (B) = 360 and (AB) = 35. Prepare 2 × 2 contingency table and compute Yule' s co-efficient of association and interpret the result.

23. In an experiment on immunization of cattle from tuberculosis the following results were obtained.

	Affected	Unaffected
Inoculated	12	26
Not inoculated	16	6

By calculating Yule' s co-efficient of association, examine the effect of vaccine is in controlling the disease.

24. Calculate the co-efficient of association between the intelligence of fathers and sons from the following data

Intelligent fathers with intelligent sons	=	300
Intelligent fathers with dull sons	=	100
Dull fathers with intelligent sons	=	50
Dull fathers with dull sons	=	500

25. Out of 3000 unskilled workers of a factory, 2000 come from rural area and out of 1200 skilled workers 300 come from rural area. Determine the association between skill and residence

26. In an anti-malarial campaign in a certain area, quinine was administrated to 812 persons out of a total population of 3428. The number of fever cases is shown below:

Treatment	Fever	No Fever
Quinine	20	792
No quinine	220	2216

Examine the effect of quinine on controlling malaria.

27. 1500 candidates appeared for competitive examinations 425 were successful. 250 had attended a coaching class and of

these 150 came out successful. Estimate the utility of the coaching class.

28. In an examination at which 600 candidates appeared of them 348 were boys. Number of passed candidates exceeded the number of failed candidates by 310. Boys failing in the examination numbered 88. Find the co-efficient of association between male sex and success in examination.

29. Following data relate to literacy and unemployment in a group of 500 persons. Calculate Yule' s co-efficient of association between literacy and unemployment and interpret it

Literate unemployed	=	220
Literate employed	=	20
Illiterate Employed	=	180

30. In a group of 400 students, the number of married is 160. Out of 120 students who failed 48 belonged to the married group. Find out whether the attributes of marriage and failure are independent.

Answers

I.

1. (a) 2. (b) 3. (d) 4. (b) 5. (b)

II.

6. Dichotomy 7. Negative 8. 0

9. $AB < \frac{(A).(B)}{N}$ 10. Inconsistent

IV.

16

	A	α	Total
B	35	30	65
β	20	15	35
Total	55	45	100

17.

	A	α	Total
B	9	4	13
β	14	37	51
Total	23	41	64

Total No of observations = 64

18. Inconsistent

19. A and B are independent

20. Inconsistent

21. 0.167

22. – 0.606, Negative association

23. – 0.705, Vaccine is effective

24. + 0.935

25. Negative association between skill and residence.

26. – 0.59. Negative association $\therefore$ quinine is effective.

27. + 0.68. Coaching class are useful

28. – 0.07

29. 0.92 Positive association between literacy and unemployment

30. Q = 0, Marriage and failure are independent.

10. DECISION THEORY

10. 0 Introduction:

Decision theory is primarily concerned with helping people and organizations in making decisions. It provides a meaningful conceptual frame work for important decision making. The decision making refers to the selection of an act from amongst various alternatives, the one which is judged to be the best under given circumstances.

The management has to consider phases like planning, organization, direction, command and control. While performing so many activities, the management has to face many situations from which the best choice is to be taken. This choice making is technically termed as "decision making" or decision taking. A decision is simply a selection from two or more courses of action. Decision making may be defined as - " a process of best selection from a set of alternative courses of action, that course of action which is supposed to meet objectives upto satisfaction of the decision maker.

The knowledge of statistical techniques helps to select the best action. The statistical decision theory refers to an optimal choice under condition of uncertainty. In this case probability theory has a vital role, as such, this probability theory will be used more frequently in the decision making theory under uncertainty and risk.

The statistical decision theory tries to reveal the logical structure of the problem into alternative action, states of nature, possible outcomes and likely pay-offs from each such outcome. Let us explain the concepts associated with the decision theory approach to problem solving.

The decision maker:

The decision maker refers to individual or a group of individual responsible for making the choice of an appropriate course of action amongst the available courses of action.

Acts (or courses of action):

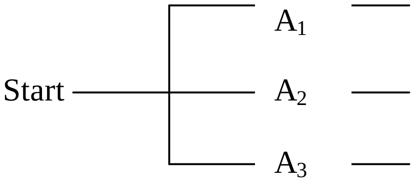
Decision making problems deals with the selection of a single act from a set of alternative acts. If two or more alternative courses of action occur in a problem, then decision making is necessary to select only one course of action.

Let the acts or action be a_1, a_2, $a_{3,}$…then the totality of all these actions is known as action space denoted by A. For three actions a_1, a_2 a_3; A = action space = (a_1, a_2, a_3) or A = (A_1, A_2, A_3). Acts may be also represented in the following matrix form i.e., either in row or column was

Acts
A_1
A_2
. .
An

Acts	A_1	A_2	…	A_n

In a tree diagram the acts or actions are shown as



Events (or States of nature):

The events identify the occurrences, which are outside of the decision maker’ s control and which determine the level of success for a given act. These events are often called ‘ States of nature’ or outcomes. An example of an event or states of nature is the level of market demand for a particular item during a stipulated time period.

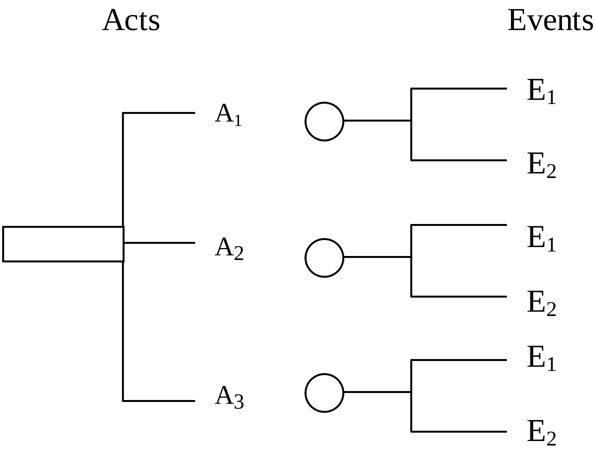
A set of states of nature may be represented in any one of the following ways:

$$S = \{S_1, S_2, . .,S_n\}$$
$$\text{or } E = \{E_1, E_2, . .,E_n\}$$
$$\text{or } \Omega = \{\theta_{1,} \theta_2, \theta_3\}$$

For example, if a washing powder is marketed, it may be highly liked by outcomes (outcome θ_1) or it may not appeal at all (outcome θ_2) or it may satisfy only a small fraction, say 25% (outcome θ_3)

$\therefore \Omega = \{ \theta_1 , \theta_2, \theta_3\}$

In a tree diagram the places are next to acts. We may also get another act on the happening of events as follows:



In matrix form, they may be represented as either of the two ways:

States of nature → ↓ Acts	S_1	S_2
A_1 A_2		

OR

Acts → States of nature ↓	A_1 $A_{2,}$. .,A_n
S_1 S_2	

10.1 Pay-off:

The result of combinations of an act with each of the states of nature is the outcome and momentary gain or loss of each such outcome is the pay-off. This means that the expression pay-off should be in quantitative form.

Pay -off may be also in terms of cost saving or time saving. In general, if there are k alternatives and n states of nature, there will be k × n outcomes or pay-offs. These k × n pay-offs can be very conveniently represented in the form of a k × n pay -off table.

States of nature	Decision alternative			
	A_1	A_2	....	A_k
E_1	a_{11}	a_{12}	....	a_{1k}
E_2	a_{21}	a_{22}	....	a_{2k}
.	.	.	....	.
.	.	.	....	.
.	.	.	....	.
E_n	a_{n1}	a_{n2}	....	a_{nk}

where a_{ij} = conditional outcome (pay-off) of the i^{th} event when j^{th} alternative is chosen. The above pay-off table is called pay-off matrix.

For example,

A farmer can raise any one of three crops on his field. The yields of each crop depend on weather conditions. We have to show pay –off in each case, if prices of the three products are as indicated in the last column of yield matrix.

Yield in kg per hectare	Weather				
		Dry (E_1)	Moderate (E_2)	Damp (E_3)	Price Rs.per . kg
	Paddy (A_1)	500	1700	4500	1.25
	Gound nut (A_2)	800	1200	1000	4.00
	Tobacco (A_3)	100	300	200	15.00

Solution:

Pay - off Table

	E_1	E_2	E_3
A_1	$500 \times 1.25 = 625$	$1700 \times 1.25 = 2125$	$4500 \times 1.25 = 5625$
A_2	$800 \times 4 = 3200$	$1200 \times 4 = 4800$	$1000 \times 4 = 4000$
A_3	$100 \times 15 = 1500$	$300 \times 15 = 4500$	$200 \times 15 = 3000$

10.1.1 Regret (or Opportunity Loss):

The difference between the highest possible profit for a state of nature and the actual profit obtained for the particular action taken is known as opportunity loss. That is an opportunity loss is the loss incurred due to failure of not adopting the best possible course of action. Opportunity losses are calculated separately for each state of nature. For a given state of nature the opportunity loss of possible course of action is the difference between the pay-off value for that course of action and the pay-off for the best possible course of action that could have been selected.

Let the pay-off of the outcomes in the 1st row be P_{11}, P_{12}……P_{1n} and similarly for the other rows.

Pay-off table

	States of nature		
Acts	S_1	S_2	S_n
A_1	P_{11}	P_{12}	P_{1n}
A_2	P_{21}	P_{22}	P_{2n}
.	.	.	.
.	.	.	.
.	.	.	.
A_m	P_{m1}	P_{m2}	P_{mn}

Consider a fixed state of nature Si. The pay-off corresponding to the n strategies are given by $P_{i1,}$ $P_{i2,}$. .,$P_{in.}$ Suppose M_i is the maximum of these quantities. The P_{i1} if A_1 is used by the decision maker there is loss of opportunity of $M_1 – P_{i1}$, and so on

Then a table showing opportunity loss can be computed as follows:

Regret (or opportunity loss table)

	States of nature		
Acts	S_1	S_2 …	S_n
A_1	M_1- P_{11}	M_2 - P_{12} …	Mn - P_{1n}
A_2	M_1- P_{21}	M_2 - P_{22} …	M_n - P_{2n}
.	.	.	.
.	.	.	.
.	.	.	.
A_m	M_1-P_{m1}	M_2 - P_{m2} …	M_n - P_{mn}

Types of decision making:

Decisions are made based upon the information data available about the occurrence of events as well as the decision situation. There are three types of decision making situations: certainty , uncertainty and risk.

Decision making under certainty:

In this case the decision maker has the complete knowledge of consequence of every decision choice with certainty. In this decision model, assumed certainty means that only one possible state of nature exists.

Example 1:

A canteen prepares a food at a total average cost of Rs. 4 per plate and sells it at a price of Rs 6. The food is prepared in the morning and is sold during the same day. Unsold food during the same day is spoiled and is to be thrown away. According to the past sale, number of plates prepared is not less than 50 or greater than 53. You are to formulate the (i) action space (ii) states of nature space (iii) pay-off table (iv) loss table

Solution:

(i) The canteen will not prepare less than 50 plates or more than 53 plates. Thus the acts or courses of action open to him are

a_1 = prepare 50 plates
a_2 = prepare 51 plates
a_3 = prepare 52 plates
a_4 = prepare 53 plates

Thus the action space is

$$A = \{a_1, a_2, a_3, a_4\}$$

(ii) The state of nature is daily demand for food plates. Then are four possible state of nature ie

S_1 = demand is 50 plates
S_2 = demand is 51 plates
S_3 = demand is 52 plates
S_4 = demand is 53 plates

Hence the state of nature space, $S = \{S_1, S_2, S_3, S_4\}$

iii) The uncertainty element in the given problem is the daily demand. The profit of the canteen is subject to the daily demand.

Let n = quantity demanded
m = quantity produced

For $n \geq m$, profit = (Cost price – Selling price) x m
= (6 – 4) x m = 2m

For $m > n$,

profit ={(Cost price – Selling price) x n}- {Cost price x (m-n) }
= 2n – 4 (m-n) = 6n – 4m

Pay-off table

Supply (m)	Demand (n) (S_1) 50	 (S_2) 51	 (S_3) 52	 (S_4) 53
(a_1) 50	100	100	100	100
(a_2) 51	96	102	102	102
(a_3) 52	92	98	104	104
(a_4) 53	88	94	100	106

(iv) To calculate the opportunity loss we first determine the maximum pay-off in each state of nature. In this state

First maximum pay-off = 100
Second maximum pay-off = 102
Third maximum pay-off = 104
Fourth maximum pay-off = 106

Loss table corresponding to the above pay-off table

Supply (m)	Demand (n) (S_1) 50	 (S_2) 51	 (S_3) 52	 (S_4) 53
(a_1) 50	100 -100 = 0	102-100 = 2	104-100 = 4	106 -100 = 6
(a_2) 51	100 - 96 = 4	102-102 = 0	104-102 = 2	106 -102 = 4
(a_3) 52	100 - 92 = 8	102 - 98 = 4	104-104 = 0	106 -104 = 2
(a_4) 53	100 - 88 =12	102 - 94 = 8	104-100 = 4	106 -106 = 0

10.2 Decision making under uncertainty (without probability):

Under conditions of uncertainty, only pay-offs are known and nothing is known about the lilkelihood of each state of nature. Such situations arise when a new product is introduced in the market or a new plant is set up. The number of different decision criteria available under the condition of uncertainty is given below.

Certain of optimism (Maximax):

The maximax criterion finds the course of action or alternative strategy that maximizes the maximum pay-off. Since this decision criterion locates the alternative with the highest possible gain, it has also been called an optimistic decision criterion. The working method is

(i) Determine the best outcome for each alternative.

(ii) Select the alternative associated with the best of these.

Expected Monetary value (EMV):

The expected monetary value is widely used to evaluate the alternative course of action (or act). The EMV for given course of action is just sum of possible pay-off of the alternative each weighted by the probability of that pay-off occurring.

The criteria of pessimism or Maximin:

This criterion is the decision to take the course of action which maximizes the minimum possible pay-off. Since this decision criterion locates the alternative strategy that has the least possible loss, it is also known as a pessimistic decision criterion. The working method is:

1) Determine the lowest outcome for each alternative.
2) Choose the alternative associated with the best of these.

Minimax Regret Criterion (Savage Criterion):

This criterion is also known as opportunity loss decision criterion because decision maker feels regret after adopting a wrong course of action (or alternative) resulting in an opportunity loss of pay-off. Thus he always intends to minimize this regret. The working method is

(a) Form the given pay-off matrix, develop an opportunity loss (or regret) matrix.

   (i) find the best pay-off corresponding to each state of nature and

   (ii) subtract all other entries (pay-off values) in that row from this value.

(b) Identify the maximum opportunity loss for each alternatives.

(c) Select the alternative associated with the lowest of these.

Equally likely decision (Baye' s or Laplace)Criterion:

Since the probabilities of states of nature are not known, it is assumed that all states of nature will occur with equal probability. ie., each state of nature is assigned an equal probability. As states of nature are mutually exclusive and collectively exhaustive, so the probability of each these must be 1 /(number of states of nature). The working method is

(a) Assign equal probability value to each state of nature by using the formula:
1/(number of states of nature)

(b) Compute the expected (or average) value for each alternative by multiplying each outcome by its probability and then summing.

(c) Select the best expected pay-off value (maximum for profit and minimum for loss)

This criterion is also known as the criterion of insufficient reason because, expect in a few cases, some information of the likelihood of occurrence of states of nature is available.

Criterion of Realism (Hurwicz Criterion):

This criterion is a compromise between an optimistic and pessimistic decision criterion. To start with a co-efficient of optimism α ($0 \leq \alpha \leq 1$) is selected.

When α is close to one, the decision maker is optimistic about the future and when α is close to zero, the decision maker is pessimistic about the future.

According to Hurwicz , select strategy which maximizes H = α (maximum pay-off in row) + ($1 - \alpha$) minimum pay-off in row.

Example 2:

Consider the following pay-off (profit) matrix

Action	States			
	(S_1)	(S_2)	(S_3)	(S_4)
A_1	5	10	18	25
A_2	8	7	8	23
A_3	21	18	12	21
A_4	30	22	19	15

No Probabilities are known for the occurrence of the nature states . Compare the solutions obtained by each of the following criteria:
(i) Maximin (ii) Laplace (iii) Hurwicz (assume that $\alpha = 0.5$)

Solution:

i) Maximin Criterion:

					Minimum
A_1:	5	10	18	25	5
A_2:	8	7	8	23	7
A_3:	21	18	12	21	12
A_4:	30	22	19	15	**15 maximum**

Best action is A_4

ii) Laplace criterion

$E(A_1) = 1/4\ [5 + 10 + 18 + 25] = 14.5$
$E(A_2) = 1/4\ [8 + 7 + 8 + 23] = 11.5$
$E(A_3) = 1/4\ [21 + 18 + 12 + 21] = 18.0$
$E(A_4) = 1/4\ [30 + 22 + 19 + 15] =$ **21.5 maximum**

$E(A_4)$ is maximum. So the best action is A_4

iv) Hurwicz Criterion (with $\alpha = 0.5$)

	Minimum	Maximum	α (max) + $(1-\alpha)$ min
A_1	5	25	0.5(25) + 0.5(5) = 15
A_2	7	23	0.5(7) + 0.5 (23) = 15
A_3	12	21	0.5(12) + 0.5 (21) = 16.5
A_4	15	30	0.5(15) + 0.5 (30) = **22.5 maximum**

Best action is A_4

Example 3:

Suppose that a decision maker faced with three decision alternatives and two states of nature. Apply (i) Maximin and (ii) Minimax regret approach to the following pay-off table to recommend the decisions.

States of Nature Act	S_1	S_2
A_1	10	15
A_2	20	12
A_3	30	11

Solution:

(i) Maximin

Act	Minimum
A_1	10
A_2	**12 maximum**
A_3	11

Act A_2 is recommended

ii) Minimax regret

States of Nature Act	S_1	S_2	Maximum Regret
A_1	30-10 = 20	15-15 = 0	20
A_2	30-20 = 10	15-12 = 3	10
A_3	30-30 = 0	15-11 = 4	**4**

Minimum of the maximum regrets is 4 which corresponds to the act A_3. So the act A_3 is recommended

Example 4:

A business man has to select three alternatives open to him each of which can be followed by any of the four possible events. The conditional pay-off (in Rs) for each action event combination are given below:

Alternative	Pay-offs conditional events			
	A	B	C	D
X	8	0	-10	6
Y	- 4	12	18	- 2
Z	14	6	0	8

Determine which alternative should the businessman choose, if he adopts the

a) Maximin criterion
b) Maximax criterion
c) Hurwicz criterion with degree of optimism is 0.7
d) Minimax regret Criterion
e) Laplace criterion

Solution:

For the given pay-off martrix, the maximum assured and minimum possible pay-off for each alternative are as given below.

Alternative	Maximum pay-off (Rs)	Minimum pay-off (Rs)	(α =0.7) H = α (maximum pay-off) + (1- α)(minimum pay-off)
X	8	-10	2.6
Y	18	- 4	**11.4**
Z	14	0	9.8

a) Since z yields the maximum of the minimum pay-off, under maximin criterion, alternative z would be chosen.
b) Under maximax criterion, the businessman would choose the alternative Y.
c) It will be optimal to choose Y under Hurwicz Criterion.
d) For the given pay-off matrix, we determine the regrets as shown below, when the regret payoffs amounts when event A occurs, are computed by the relation

Regret pay-off = maximum pay-offs from A – pay-off. Similarly for the other events.

Alternative	Pay- off amount				Regret pay-off amount				Maximum Regret
	A	B	C	D	A	B	C	D	
X	8	0	-10	6	6	12	28	2	28
Y	- 4	12	18	- 2	18	0	0	10	18
Z	14	6	0	8	0	6	18	0	18
Maximum pay-off	14	12	18	8					

Since alternative Y and Z both corresponding to the minimal of the maximum possible regrets, the decision maker would choose either of these two

(e) Laplace Criterion

In this method assigning equal probabilities to the pay-off of each strategy, results in the following expected pay-off.

Alternative	Pay-off				Expected pay-off value
	A P=1/4	B P=1/4	C P=1/4	D P=1/4	
X	8	0	–10	6	¼[8 + 0 –10 + 6] = 1
Y	- 4	12	18	- 2	¼[- 4 + 12 +18 -2] = 6
Z	14	6	0	8	¼[14 + 6 +0 + 8] = 7

Since the expected pay-off value for z is the maximum the businessman would choose alternative z

10.3 Decision making under risk (with probability):

Here the decision maker faces many states of nature. As such, he is supposed to believe authentic information, knowledge, past experience or happenings to enable him to assign probability values to the likelihood of occurrence of each state of nature. Sometimes with reference to past records, experience or information, probabilities to future events could be allotted. On the basis of probability distribution of the states of nature, one may select the best course of action having the highest expected pay-off value.

Example 5:

The pay-off table for three courses of action (A) with three states of nature (E) (or events) with their respective probabilities (p) is given. Find the best course of action.

Events	E_1	E_2	E_3
Probability → Acts ↓	0.2	0.5	0.3
A_1	2	1	-1
A_2	3	2	0
A_3	4	2	1

The expected value for each act is

A_1 : $2(0.2) + 1(0.5) - 1(0.3) = 0.6$

A_2 : $3(0.2) + 2(0.5) + 0(0.3) = 1.6$

A_3 : $4(0.2) + 2(0.5) + 1(0.3) = 2.1$

The expected monetary value for the act 3 is maximum. Therefore the best course of action is A_3.

Example 6:

Given the following pay-off of 3 acts: A_1, A_2, A_3 and their events E_1, E_2, E_3.

Act States of Nature	A_1	A_2	A_3
E_1	35	-10	-150
E_2	200	240	200
E_3	550	640	750

The probabilities of the states of nature are respectively 0.3, 0.4 and 0.3. Calculate and tabulate EMV and conclude which of the acts can be chosen as the best.

Solution:

Events	Prob.	A_1	A_2	A_3
E_1	0.3	$35 \times 0.3 = 10.5$	$-10 \times 0.3 = -3$	$-150 \times 0.3 = -45$
E_2	0.4	$200 \times 0.4 = 80.0$	$240 \times 0.4 = 96$	$200 \times 0.4 = 80$
E_3	0.3	$550 \times 0.3 = 165.0$	$640 \times 0.3 = 192$	$750 \times 0.3 = 225$
EMV		255.5	285	260

The EMV of A_2 is maximum, therefore to choose A_2

Example 7:

A shop keeper has the facility to store a large number of perishable items. He buys them at a rate of Rs.3 per item and sells at the rate of Rs.5 per item. If an item is not sold at the end of the day then there is a loss of Rs.3 per item. The daily demand has the following probability distribution.

Number of Items demanded	3	4	5	6
Probability :	0.2	0.3	0.3	0.2

How many items should he stored so that his daily expected profit is maximum?

Solution:

Let m = number of items stocked daily

n = number of items demanded daily

Now, for $n \geq m$, profit = 2m

And for $m > n$, profit = $2n - 3(m-n)$

$= 2n - 3m + 3n = 5n - 3m$

Pay - off table

Stock (m)	**Demand (n)**			
	3	**4**	**5**	**6**
3	6	6	6	6
4	3	8	8	8
5	0	5	10	10
6	–3	2	7	12
Probability	0.2	0.3	0.3	0.2

Stock(m)	Expected gain
3	$6 \times 0.2 + 6 \times 0.3 + 6 \times 0.3 + 6 \times 0.2$ = Rs. 6.00
4	$3 \times 0.2 + 8 \times 0.3 + 8 \times 0.3 + 8 \times 0.2$ = Rs. 7.00
5	$0 \times 0.2 + 5 \times 0.3 + 10 \times 0.3 + 10 \times 0.2$ = Rs. 6.50
6	$-3 \times 0.2 + 2 \times 0.3 + 7 \times 0.3 + 12 \times 0.2$ = Rs. 4.50

Thus the highest expected gain is Rs 7.00 when 4 units stocked. So, he can store 4 items to get maximum expected profit daily.

Example 8:

A magazine distributor assigns probabilities to the demand for a magazine as follows:

Copies demanded :	2	3	4	5
Probability :	0.4	0.3	0.2	0.1

A copy of magazine which he sells at Rs.8 costs Rs6. How many should he stock to get the maximum possible expected profit if the distributor can return back unsold copies for Rs.5 each?

Solution:

Let m = no of magazines stocked daily
n = no of magazines demanded

Now,

For $n \geq m$, profit = Rs 2m

and for $m > n$, profit $= 8n - 6m + 5(m-n)$

$= 8n - 6m + 5m - 5n$

$= 3n - m$

Pay-off table

Stock (m)	**Demand (n)**			
	2	**3**	**4**	**5**
2	4	4	4	4
3	3	6	6	6
4	2	5	8	8
5	1	4	7	10

Probability	0.4	0.3	0.2	0.1
Stock	**Expected Profit (in Rs)**			
2	$4 \times 0.4 + 4 \times 0.3 + 4 \times 0.2 + 4 \times 0.1 = 4.0$			
3	$3 \times 0.4 + 6 \times 0.3 + 6 \times 0.2 + 6 \times 0.1 = 4.8$			
4	$2 \times 0.4 + 5 \times 0.3 + 8 \times 0.2 + 8 \times 0.1 = 4.7$			
5	$1 \times 0.4 + 4 \times 0.3 + 7 \times 0.2 + 10 \times 0.1= 4.0$			

Thus the highest expected profit is Rs. 4.8, when 3 magazines stocked. So, the distributor can stock 3 magazines to get the maximum possible expected profit.

10.4 Decision Tree Analysis:

A decision problem may also be represented with the help of a diagram. It shows all the possible courses of action, states of nature, and the probabilities associated with the states of nature. The ‘ decision diagram’ looks very much like a drawing of a tree, therefore also called ‘ decision tree’ .

A decision tree consists of nodes, branches, probability estimates and pay-offs. Nodes are of two types, decision node (designated as a square) and chance node (designated as a circle). Alternative courses of action originate from decision node as the main branches (decision branches). Now at the terminal point of decision node, chance node exists from where chance nodes, emanate as sub-branches. The respective pay-offs and the probabilities associated with alternative courses, and the chance events are shown alongside the chance branches. At the terminal of the chance branches are shown the expected pay-off values of the outcome.

There are basically two types of decision trees-deterministic and probabilistic. These can further be divided into single stage and multistage trees. A single stage deterministic decision tree involves making only one decision under conditions of certainty (no chance events). In a multistage deterministic tree a sequence or chain of decisions are to be made, The optimal path (strategy) is one that corresponds to the maximum EMV.

In drawing a decision tree, one must follow certain basic rules and conventions as stated below:

1. Identify all decisions (and their alternatives) to be made and the order in which they must be made.
2. Identify the chance events or state of nature that might occur as a result of each decision alternative.
3. Develop a tree diagram showing the sequence of decisions and chance events. The tree is constructed starting from left and moving towards right. The square box denotes a decision point at which the available courses of action are considered. The circle O represents the chance node or event, the various states of nature or outcomes emanate from this chance event.
4. Estimate probabilities that possible events or states of nature will occur as a result of the decision alternatives.
5. Obtain outcomes (usually expressed in economic terms) of the possible interactions among decision alternatives and events.
6. Calculate the expected value of all possible decision alternatives.
7. Select the decision alternative (or course of action) offering the most attractive expected value

Advantages of decision tree:

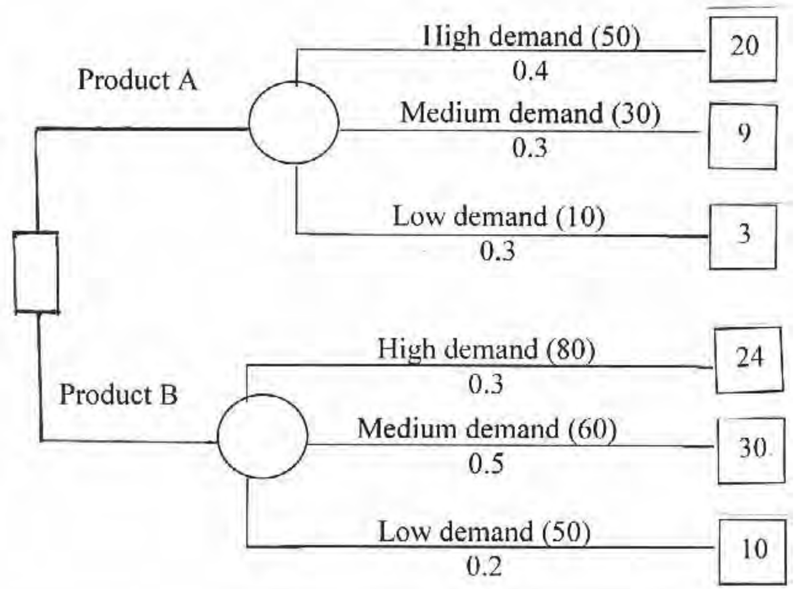
1. By drawing of decision tree, the decision maker will be in a position to visualise the entire complex of the problem.
2. Enable the decision - maker to see the various elements of his problem in content and in a systematic way.
3. Multi-dimensional decision sequences can be strung on a decision tree without conceptual difficulties.
4. Decision tree model can be applied in various fields such as introduction of a new product, marketing strategy etc…

Example 9:

A manufacturing company has to select one of the two products A or B for manufacturing. Product A requires investment of Rs.20,000 and product B Rs 40,000. Market research survey shows high, medium and low demands with corresponding probabilities and returns from sales in Rs. Thousand for the two products in the following table.

Market demand	Probability		Return from sales	
	A	B	A	B
High	0.4	0.3	50	80
Medium	0.3	0.5	30	60
Low	0.3	0.2	10	50

Construct an appropriate decision tree. What decision the company should take?



Market demand	A			B		
	X(' 000)	P	PX	X(' 000)	P	PX
High	50	0.4	20	80	0.3	24
Medium	30	0.3	9	60	0.5	30
Low	10	0.3	3	50	0.2	10
Total			32			64

Product	Return (Rs)	Investment(Rs)	Profit (Rs)
A	32,000	20,000	12,000
B	64,000	40,000	24,000

Since the profit is high in case of product B, so the company' s decision in favour of B.

Example 10:

A farm owner is considering drilling a farm well. In the past, only 70% of wells drilled were successful at 20 metres of depth in that area. Moreover, on finding no water at 20 metres, some person drilled it further up to 25 metres but only 20% struck water at 25 metres. The prevailing cost of drilling is Rs.500 per metres. The farm owner has estimated that in case he does not get his own well, he will have to pay Rs.15000 over the next 10 years to buy water from the neighbour.

Draw an appropriate decision tree and determine the farm owner' s strategy under EMV approach.

Solution:

The given data is represented by the following decision tree diagram.

Decision	Event	Probability	Cash out flows	Expected cash out flow
Decision at point D_2				
1. Drill upto 25 metres	Water struck	0.2	Rs.12500	Rs 2500
	No water struck	0.8	Rs27500	Rs 22000
			EMV(out flows)	Rs.24500

2. Do not drill	EMV (out flow) = Rs.25000

The decision at D_2 is : Drill up to 25 metres

Decision at point D_1				
1. Drill upto 20 metres	Water struck	0.7	Rs.10000	Rs 7000
	No water struck	0.3	Rs.24500	Rs.7350
			EMV(out flows)	Rs.14,350
2. Do not drill	EMV (out flow) = Rs.15000			

The decision at D_1 is : Drill up to 20 metres.
Thus the optimal strategy for the farm-owner is to drill the well up to 20 metres.

Exercise – 10

I. Choose the correct answers:

1. Decision theory is concerned with
   (a) The amount of information that is available
   (b) Criteria for measuring the ‘ goodness’ of a decision
   (c) Selecting optimal decisions in sequential problems
   (d) All of the above
2. Which of the following criteria does not apply to decision – making under uncertainly
   (a) Maximin return
   (b) Maximax return
   (c) Minimax return
   (d) Maximize expected return
3. Maximin return, maximax return and minimax regret are criteria that
   (a) Lead to the same optimal decision.
   (b) Cannot be used with probabilities
   (c) Both a and b
   (d) None of the above

4. Which of the following does not apply to a decision tree?
   (a) A square node is a point at which a decision must be made.
   (b) A circular node represents an encounter with uncertainty.
   (c) One chooses a sequence of decisions which have the greatest probability of success.
   (d) One attempts to maximize expected return.
5. The criterion which selects the action for which maximum pay-off is lowest is known as
   (a) Max-min criterion
   (b) Min-max criterion
   (c) Max –max criterion
   (d) None of these

II. Fill in the blanks:

6. Decision trees involve _____________ of decisions and random outcomes.
7. One way to deal with decision making in the ‘ uncertainity’ context is to treat all states of nature as ____________ and maximize expected return.
8. Maximizing expected net rupee return always yields the same optimal policy as ____________ expected regret.
9. The different criteria for making decisions under risk always yields the same __________ choice.
10. In decision under uncertainty, the Laplace criterion is the least conservative while the ____________ criterion is the most conservative.

III. Answer the following:

11. Explain the meaning of ‘ statistical decision theory’
12. What techniques are used to solve decision making problems under uncertainty?
13. Write a note on decision tree.
14. What is a pay-off matrix?
15. Describe how you would determine the best decision using the EMV criterion with a decision tree.

IV. Problems:

16. The pay-off table for three courses of action (A) with three states of nature (E) (or events) with their respective probabilities (P) are given. Find the best course of action.

Events Acts	E_1	E_2	E_3
A_1	2.5	2.0	–1
A_2	4.0	2.6	0
A_3	3.0	1.8	1
Probability	0.2	0.6	0.2

17. Calculate EMV and thus select the best act for the following pay-off table:

States of nature	Probability	Pay-off (Rs) by the player		
		A	B	C
X	0.3	–2	–5	20
Y	0.4	20	–10	–5
Z	0.3	40	60	30

18. Consider the pay-off matrix

States of nature	Probability	Act A_1 do not expand	Act A_2 Expand 200 units	Act A3 Expand 400 units
High demand	0.4	2500	3500	5000
Medium demand	0.4	2500	3500	2500
Low demand	0.2	2500	1500	1000

Using EMV criterion decide the best act.

19. Apply (i) maximin (ii) minimax regret to the following pay-off matrix to recommended the decisions without any knowledge of probability.

States of nature

Act	S_1	S_2	S_3
a_1	14	8	10
a_2	11	10	7
a_3	9	12	13

20. A shop keeper of some highly perishable type of fruits sees that the daily demand X of this fruit in his area the following probability distribution.

Daily Demand (in Dozen)	:	6	7	8	9
Probability	:	0.1	0.3	0.4	0.2

He sells for Rs10.00 a dozen while he buys each dozen at Rs4.00. Unsold fruits in a day are traded on the next day at Rs.2.00 per dozen, assuming that the stocks the fruits in dozen, how many should he stock so that his expected profit will be maximum?

[Hint: profit $= 6m$ for $n \geq m$

$= 10n - 4m + 2(m-n)$

$= 8n - 2m$ for $n < m$]

21. A florist, in order to satisfy the needs of a number of regular and sophisticated customers, stocks a highly perishable flowers. A dozen flowers cost Rs 3 and sell at Rs10.00 Any flower not sold on the day are worthless. Demand distribution in dozen of flowers is as follows:

Demand	1	2	3	4
Probability	0.2	0.3	0.3	0.2

How many flowers should he stock daily in order to maximize his expected net profit?

22. A florist stock highly perishable flower. A dozen of flower costs Rs3.00 and sells for Rs.10.00 Any flower not sold the day are worthless. Demand in dozen of flowers is as follows:

Demand in dozen	0	1	2	3	4
Probability	0.1	0.2	0.4	0.2	0.1

Assuming that failure to satisfy any one customer' s request will result in future lost profit amounting to Rs.5.00, in addition to the lost profit on the immediate sale, how many flowers should the florist stock to expect maximum profit?

23. A newspaper agent' s experience shows that the daily demand x of newspaper in his area has the following probability distribution

Daily Demand(x)	300	400	500	600	700
Probability	0.1	0.3	0.4	0.1	0.1

He sales the newspapers for Rs.2.00 each while he buys each at Rs.1.00. Unsold copies are treated as scrap and each such copy fetches 10 paisa. Assuming that he stocks the news papers in multiple of 100 only. How many should he stock so that his expected profit is maximum?

24. Suppose that a decision maker faced with three decision alternatives and four states of nature. Given the following profit pay-off table.

	States of nature			
Acts	S_1	S_2	S_3	S_4
a_1	16	10	12	7
a_2	13	12	9	9
a_3	11	14	15	14

Assuming that he has no knowledge of the probabilities of occurrence of the states of nature, find the decisions to be recommended under each of the following criteria.

(i) maximin (ii) maximax (iii) minimax Regret

25. Pay-off of three acts A, B and C and states of nature X, Y and Z are given below

States of nature	Pay-off (in Rs) Acts A	 B	 C
X	– 20	–50	2000
Y	200	–100	–50
Z	400	600	300

The probabilities of the states of nature are 0.3, 0.4 and 0.3. Calculate the EMV for the above and select the best art.

Answers:

I.

1. (d) 2. (d) 3. (b) 4. (c) 5. (a)

II.

6. Sequence 7. equally likely 8. minimizing

9. Optimal 10. Minimax

IV.

16. A_2 is the best

17. Select A with the highest EMV Rs.194

18. EMV: 3200, decide, Act A_3, expand 400 units

19. (i) maximin : Act a_3

(ii) minimax regret Act a_1

20. So the shop keeper should stock 8 dozen of fruits to get maximum expected profit.

21. He should stock 3 dozen of flowers to get maximum expected net profit.

22. He stocks 3 dozen of flowers to expect maximum profit Rs.9.50

23. To stock 405 copies so that his expected profit is maximum

24. (i) Act a_3 is recommended

(ii) Act a_1 is recommended

(iii) Act a_3 is recommended

25. EMA for A is highest. So the best act is A is selected