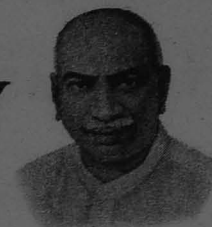




MADURAI KAMARAJ UNIVERSITY



(University with Potential for Excellence)

DISTANCE EDUCATION

B.Sc. SECOND YEAR MATHEMATICS

**PART III
PAPER - IV**

STATISTICS

UNITS : 6 to 10

Recognised by D.E.C.

www.mkudde.org

R 85

ACC-MK0
01/80.

STAR\1000 COPIES\APR'12

Unit - 6

6.1 Probability

History of Probability Theory:

A French nobleman named Antoine Gombauld (1607-1664) questioned mathematical basis for success and failure at the dice tables. He asked the French mathematician Blaise Pascal (1623-1692) "What are the odds of rolling two sixes atleast once in twenty four rolls of a pair of dice?" Pascal solved the problem having become as interested in this idea of probabilities as was Gombauld. They shared their ideas with the famous mathematician Pierre De Fermat (1601-1665) and the letters written by these three constitute the first academic journals in probability theory.

Need for Probability Theory:

Probability theory was successfully applied at the gambling tables and more relevant to our study, eventually to business. Today the mathematical theory of probability is the basis for statistical applications in both social and managerial research.

Uses of Probability Theory:

Probability is a part of our everyday life. When we hear a weather forecast of a seventy per-cent chance of rain, we change our plans from a picnic to a pool game, playing bridge. We make some probability estimate before taking decision. Managers who deal with inventories of highly styled women's clothing must wonder about the chances that sales will reach or exceed a certain level, and the buyer who stocks up on skate boards considers the probability of the life of this particular fad. In general, theory of probability is a branch of applied mathematics dealing with the effects of chance.

Here we encounter the word "chance" which is often used in everyday language but with rather indefinite meaning.

Examples:

1. A die is cubic in shape and its faces are marked from one to six. If a homogeneous good die is thrown upon a board we are certain that one of the six faces will turn up. But whether a particular face will show depends on what we call chance and cannot be predicted.
2. If a number of white and black balls, which do not differ in any respect, except in colour are concealed in an urn and we draw one of them blindly, it is certain that its colour will be either white or black, but whether it will be black or white we cannot predict that depends on chance.

Definitions

Trial and Event:

Consider an experiment which does not give unique results though repeated under essentially identical conditions, but may result in any one of the several possible outcomes. The experiment is known as a TRIAL and outcomes are known as EVENTS or CASES.

Examples:

1. Throwing of a die is a trial and getting 1 or 2 or 3 or 4 or 5 or 6 is an event.
2. Tossing of a coin is a trial and getting Head or Tail is an event.
3. Drawing three cards from a pack of well-shuffled cards is a trial and getting three aces is an event.

Exhaustive Events

All possible outcomes in any trial are known as exhaustive events or exhaustive cases.

Examples:

1. In tossing an unbiased coin once, head and tail are the two exhaustive cases.
2. In throwing a good die there are 6 exhaustive cases namely getting faces 1, 2, 3, 4, 5 and 6.
3. In drawing 4 cards from a pack of well shuffled cards, the number of exhaustive events is ${}^{52}C_4$.
4. In tossing 2 coins, the exhaustive number of cases = 2^2 namely (HH, HT, TH, TT)
5. In selecting a statistics book from a library which contains 50 statistics books, the exhaustive events are 50.

Sample Space:

The set of all exhaustive events of an experiment is known as the sample space corresponding to that experiment.

If the number of outcomes is finite, the sample space is a finite space. Otherwise, it is an infinite space. If the outcomes are discrete, the space is called a discrete sample space. If they are continuous, the space is called a continuous sample space.

Suppose an experiment is conducted to find the number of tosses of a coin to be made to get a head, the sample space will be infinite (the outcomes being H, TH, TTH, TTTH.....) but discrete.

The temperature recorded during a fixed interval on any day is an infinite continuous sample space.

An event can be defined as a subset of the sample space. If the subset is a singleton set, the event is said to be a simple event or elementary event.

Favourable Event

The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event.

Examples:

1. In tossing a coin two times the number of favourable cases for the appearance of a head and tail is 2 [(H, T)], (T, H)].
2. In throwing 2 dice, the number of favourable cases to get the sum 4 is (1, 3), (2, 2), (3, 1) ie., 3.
3. The number of favourable cases for picking 2 white balls from an urn containing 7 white balls and 3 black balls is 7C_2 .
4. In drawing a spade from a pack of 52 cards the number of favourable cases is 13.

Mutually Exclusive events

Two or more than two events are said to be mutually exclusive if they cannot happen simultaneously in the same trial i.e., one and only one of the events can take place at a time.

Examples:

1. In tossing a coin, the events head and tail are mutually exclusive, since both head and tail cannot appear simultaneously in a tossing of a coin.
2. From a bag of white and red balls, choosing a 'white' or 'red' ball are mutually exclusive.
3. In throwing a die, getting an even number and a number > 3 are not mutually exclusive.

Equally likely Events:

Two events are considered as equally probable if, after taking into consideration all relevant evidence, one of them cannot be expected in preference to the other.

Examples:

1. In tossing an unbiased coin, appearance of head and tail are equally probable events.

2. In throwing an unbiased die all the six faces are equally probable events.
3. In tossing two coins, getting two heads and getting two tails are equally likely events.
4. But getting 2 heads and getting 'one head and one tail' are not equally likely events.

Independent events

Several events are said to be independent if the occurrence of any one of them does not affect the occurrence of any other. Otherwise the events are called dependent.

Examples:

1. In tossing an unbiased coin the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.
2. If we draw a card from a pack of well-shuffled cards and replace it before drawing the second card, the result of the second draw is independent of first draw. But, if the first card drawn is not replaced then the second draw is dependent on the first draw.

Mathematical or Classical or Priori Probability:

Suppose we toss a coin. Assume that it is an unbiased coin i.e., it is uniform and the possibility of its standing on edge may be neglected. Hence the coin may fall head upwards or tail upwards.

i.e., There are two possible equally likely ways (head and tail) one of which should necessarily happen. Hence the probability of a head upwards is $\frac{1}{2}$. This is called a priori probability. i.e., Probability determined prior to the event.

Definition:

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

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

$$p = \frac{m}{n} \quad (1)$$

The number of cases favourable to the non-happening of the event E are (n-m).

The probability that E will not happen is given by $q = \frac{n-m}{n}$

It is clear that $p \geq 0$, $q \geq 0$ and

$$p + q = \frac{m}{n} + \frac{n-m}{n} = 1$$

$$0 \leq p \leq 1 \text{ and } 0 \leq q \leq 1.$$

Note: 1. Probability 'p' of the happening of an event is known as the probability of success and probability 'q' of its non-happening is known as the probability of failure.

2. If $P(E) = 1$, E is called a certain event and if $P(E) = 0$, E is called an impossible event.

The above definition of probability depends on 2 factors:

1. The sample space is finite.
2. The events are equally likely.

When one of these factors fails, the above definition cannot be applied and we have the following definition.

A Posteriori Probability or Statistical or empirical Probability

Suppose a trial is repeated a number of times under essentially homogeneous and identical conditions. Then the limit of the ratio of the number of times the event happens to, the number of trials as the number of trials become indefinitely large is called a posteriori probability.

So if 'n' denotes the number of trials and 'm' of them are favourable to an event E, then the posteriori probability $p = \lim_{n \rightarrow \infty} \frac{m}{n}$

Example 1: Seven cards are drawn at random from a pack of 52 cards. What is the probability that 4 will be red and 3 black?

Solution: No of exhaustive events = $52C_7$.

There are 26 red and 26 black cards.

4 red cards can be drawn in $26C_4$ ways and 3 black cards can be drawn in $26C_3$ ways.

Since each one of the former cases can be associated with each of the latter cases, the total number of favourable cases = $26C_4 \times 26C_3$.

$$\text{Required Probability} = \frac{26C_4 \times 26C_3}{52C_7}$$

Example 2: 4 cards are drawn at random. What is the probability that (1) all are diamonds (2) all are of the same suit.

Solution: Exhaustive no. of events = $52C_4$.

No. of ways of getting 4 diamonds = $13C_4$.

Probability of getting all diamonds = $\frac{13C_4}{52C_4}$

If all the cards should be of the same suit, then the number of favourable events = $4 \times 13C_4$

Probability of getting all 4 cards of the same suit = $\frac{4 \times 13C_4}{52C_4}$

Example 3: An urn contains 5 white and 7 black balls. Two balls are drawn at random. What is the probability that both are black?

Solution: Total number of balls = 12.

No. of ways of drawing 2 balls = $12C_2$ = No. of exhaustive events.

No. of ways of drawing 2 black balls = $7C_2$ = No. of favourable events.

Required Probability = $\frac{7C_2}{12C_2} = \frac{7 \times 6}{12 \times 11} = \frac{7}{22}$

Example 4: A bag contains 4 red and 6 black balls. Two balls are drawn at random. What is the probability of drawing 1 red and 1 black ball?

Solution: Total number of balls = 10

No. of ways of drawing 2 balls = $10C_2$.

No. of ways of drawing 1 red ball = 4.

No. of ways of drawing 1 black ball = 6.

No. of ways of drawing 1 red and 1 black ball = $4 \times 6 = 24$

Required probability = $\frac{24}{10C_2} = \frac{24 \times 1 \times 2}{10 \times 9} = \frac{8}{15}$

Example 5: If from a lottery of 30 tickets marked 1, 2, 3,, 30 four tickets are drawn. What is the probability that those marked 1 and 2 are among them.

Solution: No. of exhaustive events = ${}^{30}C_4$.

Out of the 4 tickets drawn, two are numbered 1 and 2 and the remaining two have any of the two numbers 3 to 30.

Hence, No. of favourable events = ${}^{28}C_2$.

$$\therefore \text{Required probability} = \frac{{}^{28}C_2}{{}^{30}C_4}$$

Example 6: If the letters of the word RANDOM be arranged at random, what is the chance that there are exactly two letters between A and O?

Solution: The word RANDOM consists of 6 letters. The two letters A and O can occupy ${}^6P_2 = 6 \times 5$ positions.

$\therefore$ Exhaustive events = 30.

The number of ways in which there will be exactly two letters between A and O are done as follows.

- (i) A is the first place and O is the 4th place.
- (ii) A is the 2nd place and O is the 5th place.
- (iii) A is the 3rd place and O is the 6th place.

Since A and O can interchange their positions, favourable cases = $2 \times 3 = 6$.

$$\therefore \text{Required Probability} = \frac{6}{30} = \frac{1}{5}$$

Example 7: The sum of two positive integers is 50. Find the chance that their product does not exceed 500.

Solution: The sum of two positive integers is 50.

The exhaustive cases are 25 namely (1, 49), (2, 48),, (25, 25).

Out of these 25 pairs the product of the numbers will not exceed 500 are the following pairs:
(1, 49), (2, 48),, (13, 37)

∴ The favourable number of cases = 13.

∴ Required probability = $\frac{13}{25}$

Example 8: There are four letters and four addresses envelopes. Find the chance that all letters are not despatched in the right envelopes.

Solution: First letter can be put in any one of the 4 envelopes. 2nd letter can be put in any one of the remaining 3 envelopes etc.

∴ The total number of ways in which 4 letters can be put in envelopes = 4!

All the letters can be put in right envelopes in one way only. Hence the probability of all letters being put in right envelopes = $\frac{1}{4!} = \frac{1}{24}$

The required probability = $1 - \frac{1}{24} = \frac{23}{24}$.

Exercise:

1. Prove that in a single throw of two dice, the probability of getting the sum of 7 is $\frac{1}{6}$ and the probability of getting the sum of 10 is $\frac{1}{12}$.
2. A bag contains 3 red, 6 white, 7 blue and 8 black balls. What is the probability that two balls drawn are white and blue.
3. If the letters of the word "REGULATION" be arranged at random, what is the chance that there will be exactly 4 letters between R and E.
4. Five persons are selected from a group containing 10 men, 5 women and 6 children. Find the chance that exactly three of them will be children.
5. Two numbers are chosen at random from the set of numbers 1, 2, 3, ..., 20. Find the chance that their sum is equal to 9.
6. There are four letters and four envelopes. If the letters are placed in the envelopes at random, what is the chance that atleast one letter is in the correct envelope?

Theorems on Probability

Notations

Let $A_1, A_2, \dots, A_n$ be n events.

- 1) $P(A_i)$ = probability that the event A_i occurs
- 2) $\overline{P(A_i)}$ = probability that the event A_i does not occur.
 $\therefore P(A_i) + \overline{P(A_i)} = 1.$
- 3) $P(A_i + A_j)$ = probability of occurrence of atleast one of the events A_i, A_j or both.
- 4) $P(A_1 + A_2 + \dots + A_n)$ = probability of occurrence of atleast one of the n events.
- 5) $P(A_i A_j)$ = probability of the simultaneous occurrence of both the events A_i and A_j .
- 6) $P(A_1 A_2 \dots A_n)$ = probability of simultaneous occurrence of all events.

Definition: Two events A, B are called complementary events if $A + B$ is a certain event and AB is an impossible event i.e., if $P(A+B) = 1$ and $P(AB) = 0$.

Example: $A, \overline{A}$ are complementary events.

Theorem 1: Let A_1, A_2 be two mutually exclusive events. Then $P(A_1 + A_2) = P(A_1) + P(A_2)$.

Proof: Let the number of exhaustive events be n and let the number of events favourable to A_1 , and A_2 be m_1, m_2 respectively. Since, A_1, A_2 are mutually exclusive, the number of events favourable to either A_1 or $A_2 = m_1 + m_2$.

$$\text{Hence, } P(A_1) = \frac{m_1}{n} \quad P(A_2) = \frac{m_2}{n}$$

$$P(A_1 + A_2) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A_1) + P(A_2)$$

Corollary-1: If $A_1, A_2, \dots, A_k$ are mutually exclusive events then $P(A_1 + A_2 + \dots + A_k) = P(A_1) + P(A_2) + \dots + P(A_k)$

Proof: If $m_1, m_2, \dots, m_k$ are number of cases favourable to $A_1, A_2, \dots, A_k$ then $(m_1 + m_2 + \dots + m_k)$ is the number of case favourable to A_1 or $A_2 \dots$ or A_k .

$$\therefore P(A_1 + A_2 + \dots + A_k) = \frac{m_1 + m_2 + \dots + m_k}{n} = \frac{m_1}{n} + \frac{m_2}{n} + \dots + \frac{m_k}{n}$$

$$= P(A_1) + P(A_2) + \dots + P(A_k).$$

Corollary 2: Suppose the events $A_1, A_2, \dots, A_k$ are exhaustive and also mutually exclusive, then $P(A_1) + P(A_2) + \dots + P(A_k) = 1$.

Corollary 3: If A, B are complementary events, $P(A) + P(B) = 1$.

Theorem 2: Addition Theorem of probability

If A_1, A_2 are two events, then

$$P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1 A_2).$$

Proof: The event $(A_1 + A_2)$ occurs if and only if one of the mutually exclusive events A_1 and $A_2 - (A_1 A_2)$ occurs.

$$\text{By theorem 1, } P(A_1 + A_2) = P(A_1) + P(A_2 - A_1 A_2) \quad (1)$$

But A_2 is the sum of two mutually exclusive events $(A_1 A_2)$ and $(A_2 - A_1 A_2)$.

$$\therefore P(A_2) = P(A_1 A_2) + P(A_2 - A_1 A_2) \quad (2)$$

From (1) and (2) we get

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

Corollary 1: If A_1, A_2, A_3 are 3 events then

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

Cor 2: If A_1, A_2 are mutually exclusive, $A_1 A_2$ is an empty set and hence $P(A_1 A_2) = 0$.

Example 1: What is the probability of drawing an ace or a queen or a king in a single draw out of a well-shuffled pack of cards?

Solution: Probability of drawing an ace = $\frac{4}{52} = \frac{1}{13}$

Probability of drawing a queen = $\frac{4}{52} = \frac{1}{13}$

$$\text{Probability of drawing a king} = \frac{4}{52} = \frac{1}{13}$$

Since these three are mutually exclusive, the probability of drawing anyone of these

$$= \frac{1}{13} + \frac{1}{13} + \frac{1}{13} = \frac{3}{13}$$

Example 2: If two dice are thrown, what is the probability that the sum is (a) greater than 9, (b) neither 7 nor 11.

Solution: (a) Let A be the event of getting a sum greater than 9.

$$\text{Probability of getting a sum } 10 = \frac{3}{36} = \frac{1}{12}$$

[$\therefore$ the favourable events are (4, 6), (5, 5), (6, 4)]

$$\text{Probability of getting a sum of } 11 = \frac{2}{36} = \frac{1}{18}$$

[$\therefore$ the favourable events are (5, 6), (6, 5)]

$$\text{Probability of getting a sum of } 12 = \frac{1}{36}$$

$$\text{Required probability} = \frac{1}{12} + \frac{1}{18} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}$$

(b) Let A denote the event of getting the sum of 7 and B, the event of getting the sum of 11.

$$\therefore P(A) = \frac{6}{36} \quad [\because (4, 3), (3, 4), (5, 2), (2, 5), (6, 1), (1, 6) \text{ are the favourable events}]$$

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

$$P(B) = \frac{2}{36} = \frac{1}{18}$$

$$\text{Probability of getting 7 or 11} = P(A + B)$$

$$\therefore \text{Probability of not getting 7 or 11} = 1 - P(A + B) = 1 - [P(A) + P(B)]$$

($\therefore$ A and B are mutually exclusive)

$$= 1 - \left[\frac{1}{6} + \frac{1}{18} \right] = 1 - \left[\frac{4}{18} \right]$$

$$= 1 - \frac{2}{9} = \frac{7}{9}$$

•**Example 3:** A number is chosen at random from 1 to 50. What is the probability that the number chosen is either a multiple of 2 or 5?

Solution: Let A denote the event of choosing a number, which is multiple of 2 and let B denote the event of choosing a number which is a multiple of 5. Then A and B are not mutually exclusive since whenever a multiple of 10 is chosen, both A and B occur.

∴ No of cases favourable to A = 25 (all even numbers)

$$\therefore P(A) = \frac{25}{50} = \frac{1}{2}$$

No. of cases favourable to B = 10 (multiples of 5).

$$\therefore P(B) = \frac{10}{50} = \frac{1}{5}$$

No. of cases favourable to both A and B = 5 (multiples of 10)

$$\therefore P(AB) = \frac{5}{50} = \frac{1}{10}$$

$$\therefore P(A+B) = P(A) + P(B) - P(AB)$$

$$= \frac{1}{2} + \frac{1}{5} - \frac{1}{10} = \frac{6}{10} = \frac{3}{5}$$

6.2 - Conditional Probability:

The probability of an event changes as conditions regarding its sample space change. Let us use the symbol $P(A/B)$ and read it as conditional probability of A relative to B. By this we mean the probability of occurrence of A, having known that B has occurred. i.e., We are searching for cases favourable to A within B and find the probability. i.e., we restrict the sample space to B and find the probability of the occurrence of A relative to B.

Let us consider an example. In a bag containing 15 balls, 7 are blue and the remaining white. What is the probability of drawing a blue ball after a white ball is drawn and not replaced. As one white ball has been drawn and not replaced, the sample space consists of 14 balls and 7 of them are blue. Hence the probability of drawing a blue ball after drawing a white ball is $\frac{7}{14} = \frac{1}{2}$ i.e., if

we denote the event of drawing a blue ball by 'B' and the event of drawing a white ball by 'W', we have $P(B/W) = \frac{1}{2}$

Theorem 3:- Product Theorem

Let A_1, A_2 be two events and let $P(A_1) > 0$. Then $P(A_1 A_2) = P(A_1) P(A_2/A_1)$

Proof: Let the number of exhaustive events be n and the number of events favourable to $A_1, A_2, A_1 A_2$ be m_1, m_2, m . Since $P(A_1) > 0$, $m_1 \neq 0$.

$$P(A_1 A_2) = \frac{m}{n} = \frac{m}{m_1} \cdot \frac{m_1}{n}$$

$$P(A_1) = \frac{m_1}{n}$$

Since m is the number of events favourable to A_1 and A_2 , $\frac{m}{m_1}$ is the proportion of events favourable to A_2 among the events favourable to A_1 . Hence it denotes the conditional probability of A_2 relative to A_1 .

$$\text{i.e., } P(A_2/A_1) = \frac{m}{m_1}$$

$$P(A_1 A_2) = P(A_1) P(A_2/A_1).$$

Cor-1: If $P(A_2) > 0$, $P(A_1 A_2) = P(A_1) P(A_1/A_2)$.

Cor-2: If $P(A_1) > 0$, $P(A_2/A_1) = \frac{P(A_1 A_2)}{P(A_1)}$

Definition 1:- Suppose the probability of the occurrence of the event B is unaffected by the information that the event A has occurred. i.e., $P(B) = P(B/A)$. Then we say that B is independent of A .

Cor-3: If A and B are two independent possible events, then $P(AB) = P(A) P(B)$.

Definition 2:- If there are k events $A_1, A_2, \dots, A_k$ then they are said to be mutually independent if $P(A_1 A_2 \dots A_k) = P(A_1) P(A_2) \dots P(A_k)$

Example 1- In a group of 30 persons 20 are employed and 10 unemployed. Of 20 employed, 12 are graduates. Find the probability that an employed is a graduate.

Solution: Let A denote the event of being a graduate and E , the event of being employed.

We want to find $P(A/E)$.

From the given data, we get $P(AE) = \frac{12}{30}$.

$$P(E) = \frac{20}{30} > 0.$$

From the product theorem, we have

$$P(AE) = P(E) P(A/E)$$

$$\begin{aligned}\therefore P(A/E) &= \frac{P(AE)}{P(E)} \\ &= \frac{12}{30} \div \frac{20}{30} = \frac{12}{30} = \frac{3}{5}\end{aligned}$$

Example 2: What is the chance of getting two sixes in two rollings of a single die?

Solution: Probability of getting 6 in first rolling = $\frac{1}{6}$

Probability of getting 6 in second rolling = $\frac{1}{6}$. As the two events are independent, probability of simultaneous occurrence of both = $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

Example 3: A bag contains 6 white and 9 black balls. Three drawings of 1 ball each are made such that (i) the balls are replaced before the next drawing and (ii) the balls are not replaced before the next drawing. Find the probability that all the three are black balls.

Solution: (i) If the balls are replaced before the next drawing is made, the drawings of consecutive balls are independent and probability of each of them being black is $\frac{9}{15} = \frac{3}{5}$

So, the probability of all three being black = $\frac{3}{5} \times \frac{3}{5} \times \frac{3}{5} = \frac{27}{125}$

(ii) In this case, as the balls are not replaced before the next drawing is made the sample space for the second and the third drawing changes.

Probability of first ball is black = $\frac{9}{15} = \frac{3}{5}$

$P(\text{second ball is black} / \text{first ball is black}) = \frac{8}{14}$
 $= \frac{4}{7}$

$P(\text{third ball is black} / \text{first two are blacks}) = \frac{7}{13}$

$P(\text{all three black}) = \text{Product of these probabilities}$

$$= \frac{3}{5} \times \frac{4}{7} \times \frac{7}{13} = \frac{12}{65}$$

Example 4: Three groups of children contain 3 girls and 1 boy, 2 girls and 2 boys and 1 girl and 3 boys. One child is selected at random from each group. Show that the chance that the three selected consist of 1 girl and 2 boys is $\frac{13}{32}$

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

Group No:

	I	II	III
(i)	girl	boy	boy
(ii)	boy	girl	boy
(iii)	boy	boy	girl

By addition theorem, required probability is the sum of the probabilities of three cases. Since the probability of selecting a girl from the first group is $\frac{3}{4}$, of selecting a boy from the second is $\frac{2}{4}$ and of selecting a boy from the third is $\frac{3}{4}$ and since these events are independent, probability of

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

$$\text{Similarly, probability of (ii)} = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{3}{32}$$

$$\text{and probability of (iii)} = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{1}{32}$$

$$\therefore \text{Required probability} = \frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}$$

Exercise:

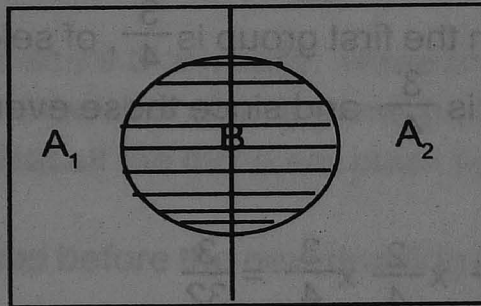
- 1) If two dice are thrown, what is the probability that the sum is (i) less than 10, (ii) neither 8 nor 12.
- 2) A bag contains 6 white and 9 black balls. Four balls are drawn at a time. Find the probability for the first draw to give 4 white and the second to give 4 black balls in each of the following cases.
 - (i) the balls are replaced before the second draw.
 - (ii) the balls are not replaced before the second draw.
- 3) If a single draw is made from a well-shuffled pack of cards what is the probability that the card is a queen or any card of hearts.
- 4) In a bag containing 15 balls, 7 are blue and the remaining white. What is the probability of drawing two balls without replacement given that the first ball is white and the second blue.

6.3: Bayes' Theorem

The concept of conditional probability 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 a specific cause. The procedure for revising these probabilities is known as Bayes' theorem.

Bayes' theorem is based on the formula for conditional probability.

Let A_1 and A_2 = The set of events which are mutually exclusive and exhaustive and B = a simple event which intersects each of the A events. i.e.,



The part of B which is within A_1 represents the area ' A_1 and B ' and the part of B within A_2 represents the area ' A_2 and B '.

Then the probability of event A_1 given event B , is $P(A_1/B) = \frac{P(A_1 \text{ and } B)}{P(B)}$ and similarly the probability of event A_2 , given B is

$$P(A_2/B) = \frac{P(A_2 \text{ and } B)}{P(B)}$$

Where $P(B) = P(A_1 \text{ and } B) + P(A_2 \text{ and } B)$

$$P(A_1 \text{ and } B) = P(A_1) \times P(B/A_1) \text{ and}$$

$$P(A_2 \text{ and } B) = P(A_2) \times P(B/A_2).$$

In general, let $A_1, A_2, A_3, \dots, A_k, \dots, A_n$ be a set of n mutually exclusive and collectively exhaustive events. If B is another event such that $P(B)$ is not zero, then

$$P(A_i/B) = \frac{P(B/A_i) P(A_i)}{\sum_{i=1}^n P(B/A_i) P(A_i)}$$

Example:- A company uses a 'selling aptitude test' in the selection of salesmen. Past experience has shown that only 70% of all persons applying for a sales position achieved a classification 'dissatisfactory' in actual selling, the remainder were classified as satisfactory. 85% had scored a passing grade on the aptitude test. Only 25% of those classified unsatisfactory had passed the test on the basis of this information. What is the probability that a candidate would be a satisfactory salesman given that he passed the aptitude test?

Solution:- If S stands for a 'satisfactory' classification as a salesman and P stands for 'passing the test' then the probability that a candidate would be 'satisfactory' salesmen given that he passed the aptitude test is

$$P(S/P) = \frac{(0.70)(0.85)}{(0.70)(0.85) + (0.30)(0.25)} = 0.888.$$

The result indicates that the tests are of value in screening candidates. Assuming no change in the type of candidates applying for the selling positions, the probability that a random applicant would be satisfactory is 70%. On the other hand, if the company only accepts an applicant if he passed the test, the probability increases to 0.888

6.4 Random Variable

A random variable is described as a variable whose values are determined by chance.

We know that the word 'variable' means that its value cannot be fixed but can take different values under different occasions. The expression "determined by chance" means that the values of the variable are determined by the outcomes of a random experiment.

Consider the random experiment of tossing 3 coins. The sample space S consists of $2^3 = 8$ points.

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

Let us consider the variable X, which is the number of heads obtained. Then X is a random variable which can take any one of the values 0, 1, 2 or 3.

Outcome	HHH	HTH	THH	TTH	HHT	HTT	THT	TTT
Value of X	3	2	2	1	2	1	1	0

We can regard X as a function with domain S and range {0, 1, 2, 3}.

Definition:- The function $X: S \rightarrow R$ is a random variable, where S is the sample space.

Distribution function:

Let X be a random variable. Then the function $F(x) = p\{X \leq x\}$, $(-\infty, < x < \infty)$ is called the distributive function of random variable x.

Value of X, x	0	1	2	3	4	5	6	7	8
$p(X=x) = p(x)$	0.125	0.375	0.375	0.125	0.0625	0.0625	0.0625	0.0625	0.0625

Properties of distribution function:

If $F(x)$ is the distribution function of the random variable x , and if $a < b$ then

1. $p(a < x < b) = F(b) - F(a)$

2. $0 \leq F(x) \leq 1$

3. If $x < y$ then $F(x) \leq F(y)$

ie., the distribution function is monotonically non decreasing and lies between 0 and 1.

4. $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ and $F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$.

Discrete random variable:

If a random variable takes atmost a countable number of values it is called a discrete random variable.

Let X be a discrete random variable taking atmost a countably infinite number of values $x_1, x_2, \dots$, with each x_i , we associate a probability $p(X = x_i) = p(x_i) = p_i$; then p_i is called the probability X which takes the value x_i , satisfying the following conditions:

(i) $p(x_i) \geq 0$ for all i .

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

This function $p(x_i)$ is called the probability mass function of the random variable X and the set $\{p(x_i)\}$ is called the probability distribution of the random variable X .

Discrete distribution function:

If $F(x)$ be the distribution function of the discrete random variable X , and the values which X takes are $x_1, x_2, \dots$, then $F(x) = \sum p_i$ where i is such that $x_i \leq x$ and $p_i \geq 0$ $\sum_{i=1}^{\infty} p_i = 1$

Example:

A random variable X has the following probability distribution:

Value of X, x	0	1	2	3	4	5	6	7	8
$p(X=x) = p(x)$	a	3a	5a	7a	9a	11a	13a	15a	17a

1. Determine the value of a

2. Find $p(x < 3)$

3. Find $p[0 < X < 5]$

4. Find $p[X \geq 3]$

5. Find the distribution function of X .

Solution:

$$1. \sum_{i=1}^{\infty} p_i = 1$$

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

$$\Rightarrow 81a = 1 \quad \therefore a = \frac{1}{81}$$

$$2. p[X < 3] = p[X=0] + p[X=1] + p[X=2]$$

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

$$= 9a = 9 \times \frac{1}{81} = \frac{1}{9}$$

$$3. p[0 < X < 5] = p[X=1] + p[X=2] + p[X=3] + p[X=4]$$

$$= 3a + 5a + 7a + 9a$$

$$= 24a = 24 \times \frac{1}{81} = \frac{8}{27}$$

$$4. p[pX \geq 3] = 1 - p[X < 3]$$

$$= 1 - \frac{1}{9} = \frac{8}{9}$$

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

The table is the distribution function of x.

Continuous Random Variable

A random variable X is said to be a continuous random variable if it takes all possible values between certain limits or in an interval which may be finite or infinite.

Example: The density of milk taken for testing at a form.

Probability density function (p.d.f)

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

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

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

Note: 1. In the case of continuous random variables the probability at a point is always zero i.e.,

$$P(x = a) = \int_a^a f(x) dx = 0$$

2. If X assumes values only in an interval (a, b) we can regard $f(x) = 0$ for all $x \notin (a, b)$.

$$\text{Hence } \int_{-\infty}^{\infty} f(x) dx = 1 \Rightarrow \int_a^b f(x) dx = 1$$

3. The probability that X lies in the interval (a, b) is given by $P(a < X < b) = \int_a^b f(x) dx$

Continuous distribution function:

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

$$(i) F(x) = \int_{-\infty}^x f(t) dt = P(X \leq x); -\infty < x < \infty$$

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

Example 1:

For the probability density function

$$f(x) = cx^2(1-x); 0 < x < 1 \text{ find}$$

$$(i) \text{ the constant } C \quad (ii) P(X < \frac{1}{2})$$

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

$$\therefore \int_0^1 cx^2(1-x)dx = 1. \text{ ie., } \int_0^1 (x^2 - x^3)dx = 1$$

$$\text{Integrating, } C \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^1 = 1$$

$$C \left(\frac{1}{3} - \frac{1}{4} \right) = 1 \Rightarrow C \left(\frac{1}{12} \right) = 1 \therefore C = 12$$

$$(ii) P(X < \frac{1}{2}) = \int_0^{\frac{1}{2}} f(x) dx = \int_0^{\frac{1}{2}} 12x^2(1-x) dx$$

$$= 12 \int_0^{\frac{1}{2}} (x^2 - x^3) dx = 12 \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^{\frac{1}{2}}$$

$$= 12 \left[\frac{(\frac{1}{2})^3}{3} - \frac{(\frac{1}{2})^4}{4} \right] = 12 \left(\frac{1}{24} - \frac{1}{64} \right)$$

$$= 12 \left(\frac{1}{8} \right) \left(\frac{1}{3} - \frac{1}{8} \right) = \frac{5}{16}$$

Example 2:- The p.d.f. of a continuous variate X is $f(x) = ke^{-|x|}$ where k is an unknown constant. Find k .

Solution:- We have $\int_{-\infty}^{\infty} f(x) dx = 1$.

$$\therefore \int_{-\infty}^{\infty} ke^{-|x|} dx = 1.$$

$$2 \int_0^{\infty} ke^{-|x|} dx = 1.$$

$$2k = 1$$

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

Exercise:

1. Examine whether $f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$; $-\infty < x < \infty$ can be a p.d.f. of a continuous random variable X .

2. A continuous random variable X has the distribution function $F(x) = \begin{cases} 0 & \text{if } x < 1 \\ k(1-x) & \text{if } 1 < x < 3 \\ 0 & \text{if } x > 3 \end{cases}$

Find (i) k (ii) the probability density function $f(x)$.

6.5 Mathematical Expectation

An average of a probability of distribution is usually called the expectation or the expected value of the random variable.

Definition: Let X be a discrete random variable which can assume any of the values $x_1, x_2, \dots, x_n$ with respective probabilities $p_1, p_2, \dots, p_n$. Then the mathematical expectation of X or the expected value of x denoted by $E(X)$ is defined as $E(X) = x_1 p_1 + x_2 p_2 + \dots + x_n p_n$.

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

Physical Interpretation of $E(X)$.

Let us consider the following frequency distribution of the random variable X .

X	x_1	x_2	x_3		x_n
f	f_1	f_2	f_3		f_n

Then the mean of the distribution is given by

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{N} \quad \text{where } N \text{ is the total frequency.}$$

$$\bar{x} = \frac{f_1}{N} x_1 + \frac{f_2}{N} x_2 + \dots + \frac{f_n}{N} x_n \quad (1)$$

Out of total N cases, f_i are favourable to x_i .

$$\therefore P(X=x_i) = \frac{f_i}{N} = p_i \text{ (say)} \quad (i=1, 2, 3, \dots, n)$$

$$\Rightarrow \frac{f_i}{N} = p_i \quad (i = 1, 2, \dots, n)$$

Substituting in (1) we get,

$$\bar{x} = p_1 x_1 + p_2 x_2 + \dots + p_n x_n$$

$$\Rightarrow \bar{x} = E(X)$$

Hence mathematical expectation of a random variable is nothing but its arithmetic mean.

Definition: Mathematical expectation of $\phi(x)$, a function of random variable X is given by

$$E(\phi(x)) = \sum p(X=x) \phi(x)$$

In particular (i) if we take $\phi(x) = x^r$ then $E(x^r) = \sum_{i=1}^n p_i x_i^r$ which is denoted as μ_r' , the r^{th} moment about origin

$$\therefore \mu_r' = E(x^r)$$

$$\therefore \mu_1' = E(x) = \text{mean} = \bar{x}$$

$$\mu_2' = E(x^2)$$

$$\text{Variance} = \mu_2 = \mu_2' - (\mu_1')^2$$

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

(ii) If we take $\phi(x) = (X - \bar{X})^r$

then $E(X - \bar{X})^r = \sum p_i (x_i - \bar{x})^r$ which is denoted as μ_r , the r^{th} moment about the mean.

$$\text{i.e., } \mu_r = E(X - \bar{X})^r = \sum p_i (x_i - \bar{x})^r$$

$$\text{when } r = 2, \mu_2 = E(X - \bar{X})^2 = \sum p_i (x_i - \bar{x})^2$$

$$\mu_2 = E[X - E(\bar{X})]^2 \quad [\because X = E(x)]$$

Addition theorem on Expectations

For two random variables X and Y , $E(X+Y) = E(X) + E(Y)$ provided $E(X)$ and $E(Y)$ exist.

ie., The mathematical expectation of the sum of random variables is equal to the sum of their expectations, provided all the expectations exist.

Multiplication theorem on Expectations

For two independent random variables X and Y , $E(XY) = E(X) \cdot E(Y)$ provided $E(X)$ and $E(Y)$ exist.

In general mathematical expectation of the product of a number of independent random variables is equal to the product of their expectations.

Theorem 1:- $E(C) = C$ where C is a constant

Proof:- $E(C) = \sum_{i=1}^n c p_i$ (by definition)

$$= C \left(\sum_{i=1}^n p_i \right)$$

$$= C \times 1 \quad \left(\because \sum_{i=1}^n p_i = 1 \right)$$

$$= C$$

Theorem 2:- $E(CX) = C E(X)$ where C is a constant.

Proof:- Let X be a random variable which takes values $x_1, x_2, \dots, x_n$ and their corresponding probabilities $p_1, p_2, \dots, p_n$.

$$E(CX) = (cx_1) p_1 + (cx_2) p_2 + \dots + (cx_n) p_n$$

$$= c(x_1 p_1 + x_2 p_2 + \dots + x_n p_n)$$

$$= C E(X)$$

Theorem 3:- $E(ax + b) = a E(X) + b$ where a and b are constants.

Proof:- Let X be a random variable which takes values $x_1, x_2, \dots, x_n$ with corresponding probabilities $p_1, p_2, \dots, p_n$.

$$E(ax + b) = p_1(ax_1 + b) + p_2(ax_2 + b) + \dots + p_n(ax_n + b)$$

$$= a(p_1x_1 + p_2x_2 + \dots + p_nx_n) + b(p_1 + p_2 + \dots + p_n)$$

$$= aE(X) + b \quad (\because \sum p_i = 1)$$

Note: Variance = $\sigma^2 = \text{Var}(X)$

$$= \frac{1}{N} \sum f(x - \bar{x})^2$$

$$= \text{Arithmetic mean of } (X - \bar{x})^2$$

$$= E(X - \bar{x})^2 \quad (\because \text{arithmetic mean is same as expected value})$$

$$\therefore \sigma^2 = E[(X - \bar{x})^2]$$

$$= E[X - E(X)]^2 \quad \text{--- (1)}$$

$$\text{Also we have } \sigma^2 = \frac{1}{N} \sum fx^2 - (\bar{x})^2$$

$$= \text{Arithmetic mean of } X^2 - (\bar{x})^2$$

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

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

(1) and (2) give the expression for variance of X in term of expectation.

Theorem 4: $\text{Var}(X + C) = \text{Var} X$ where C is a constant.

Proof: $E(X + C) = E(X) + C$

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

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

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

$$= \text{Var}(X).$$

Similarly we can prove that $\text{Var}(X - C) = \text{Var} X$. So, variance is independent of change of origin.

Theorem 5: $\text{Var}(ax) = a^2 \text{Var}(x)$

$$\begin{aligned} \text{Proof: } \text{Var}(ax) &= E[ax - E(ax)]^2 \\ &= E[ax - aE(x)]^2 \\ &= a^2 E[x - E(x)]^2 \\ &= a^2 \text{Var}(x). \end{aligned}$$

$\therefore$ Variance is not independent of change of scale.

Corollary: $\text{Var}(ax \pm b) = a^2 \text{Var} X$.

$$\begin{aligned} \text{Proof: } \text{Var}(ax \pm b) &= \text{Var}(ax) \quad (\text{by theorem 4}) \\ &= a^2 \text{Var}(x) \quad (\text{by theorem 5}) \end{aligned}$$

Theorem 6: $\text{Var}(c) = 0$ where c is a constant

$$\begin{aligned} \text{Proof: } \text{Var}(c) &= E[c - E(c)]^2 \\ &= E[c - c]^2 \quad (\because E(c) = c) \\ &= E(0) = 0 \end{aligned}$$

Example 1: A die is thrown. What is the expectation of the number on it.

Solution: Let X denote the number on the die when it is thrown. X takes values 1, 2, 3, 4, 5, 6 and has equal probability $\frac{1}{6}$.

The probability distribution is

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

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

$$\begin{aligned} &= (1 \times \frac{1}{6}) + (2 \times \frac{1}{6}) + (3 \times \frac{1}{6}) + (4 \times \frac{1}{6}) + (5 \times \frac{1}{6}) + (6 \times \frac{1}{6}) \\ &= \frac{1}{6} (1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = \frac{7}{2} \end{aligned}$$

Example 2: Let X be a discrete random variable with the following probability distribution

x	-3	6	9
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Find $E(X)$, $E(X^2)$ and $E(2X+1)^2$

Solution: $E(X) = \sum_{i=1}^n x_i P(x_i)$

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

$$= -\frac{1}{2} + 3 + 3 = \frac{11}{2}$$

$$E(X^2) = \sum_{i=1}^n x_i^2 P(x_i)$$

$$= (-3)^2 \left(\frac{1}{6}\right) + (6)^2 \left(\frac{1}{2}\right) + (9)^2 \left(\frac{1}{3}\right) = \frac{3}{2} + 18 + 27 = \frac{93}{2}$$

$$E(2X + 1)^2 = E(4X^2 + 4X + 1)$$

$$= E(4X^2) + E(4X) + E(1)$$

$$= 4E(X^2) + 4E(X) + 1$$

$$= 4 \left(\frac{93}{2}\right) + 4 \left(\frac{11}{2}\right) + 1$$

$$= 209$$

Expectation of a continuous Random variable

Definition: Let X be a continuous random variable with p.d.f., then the mathematical expectation of

X is defined as $E(X) = \int_{-\infty}^{\infty} xf(x) dx$.

Note 1: $\int_{-\infty}^{\infty} f(x) dx = 1$ provided the integral exists.

Note 2: If ϕ is a function such that $\phi(x)$ is a random variable and $E[\phi(x)]$ exists then

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

Results: (i) $E(c) = c$ where c is a constant.

$$E(c) = \int_{-\infty}^{\infty} f(x) dx = c \int_{-\infty}^{\infty} f(x) dx = c(1) = c.$$

$$(ii) E(aX \pm b) = aE(x) + b$$

$$\begin{aligned} E(aX \pm b) &= \int_{-\infty}^{\infty} (ax \pm b) f(x) dx \\ &= a \int_{-\infty}^{\infty} x f(x) dx \pm b \int_{-\infty}^{\infty} f(x) dx = aE(x) \pm b \end{aligned}$$

Example 1: For the following p.d.f. compute mean and variance $f(x)$:

$$f(x) = Ax(1-x); \text{ if } 0 \leq x \leq 1$$

$$= 0 \quad \text{otherwise}$$

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

$$\therefore \int_0^1 Ax(1-x) dx = 1$$

$$A \int_0^1 (x-x^2) dx = 1 \Rightarrow A \left(\frac{x^2}{2} - \frac{x^3}{3} \right)_0^1 = 1$$

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

$$A \left(\frac{1}{6} \right) = 1 \Rightarrow A = 6$$

$$\text{Mean} = E(x) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 6x(1-x) dx$$

$$= 6 \int_0^1 (x^2 - x^3) dx$$

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

$$E(X) = \frac{1}{2}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2 \cdot 6x(1-x) dx$$

$$= 6 \int_0^1 (x^3 - x^4) dx = 6 \left(\frac{x^4}{4} - \frac{x^5}{5} \right)_0^1$$

$$= 6 \left(\frac{1}{4} - \frac{1}{5} \right) = \frac{6}{20} = \frac{3}{10}$$

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

$$= \frac{3}{10} - \frac{1}{4} = \frac{6-5}{20} = \frac{1}{20} = 0.05$$

$$\text{Mean} = 0.5$$

$$\text{Variance} = 0.05$$

Example 2: The p.d.f. of a continuous random variable X is given by $f(x) = \frac{x}{2}$, if $0 < x < 2$
 $= 0$, otherwise.

Find (a) Expected value of X (b) $E(3x^2 - 2x)$

$$\begin{aligned}\text{Solution: (a) } E(X) &= \int_0^2 x f(x) dx \\ &= \int_0^2 x \left(\frac{x}{2}\right) dx = \frac{1}{2} \int_0^2 x^2 dx \\ &= \frac{1}{2} \left(\frac{x^3}{3}\right)_0^2 = \frac{4}{3}\end{aligned}$$

$$\begin{aligned}\text{(b) } E(3X^2 - 2X) &= \int_0^2 (3x^2 - 2x) f(x) dx \\ &= \int_0^2 (3x^2 - 2x) \left(\frac{x}{2}\right) dx \\ &= \frac{1}{2} \int_0^2 (3x^3 - 2x^2) dx \\ &= \frac{1}{2} \left[\frac{3x^4}{4} - \frac{2x^3}{3} \right]_0^2 \\ &= \frac{1}{2} \left[12 - \frac{16}{3} \right] = \frac{20}{6} = \frac{10}{3}\end{aligned}$$

Exercise

1. A random variable X has the following probability distribution

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

Find (i) $E(X)$ (ii) $\text{Var}(X)$ and (iii) $E(2X + 3)^2$

2. Find the mean and variance of the following p.d.f.s

$$\text{(a) } f(x) = \frac{1}{2} (x+1), \quad -1 < x < 1$$

$$= 0 \quad \text{otherwise}$$

$$\text{(b) } f(x) = e^{-x}, \quad \text{if } x \geq 0$$

$$= 0, \quad \text{if } x < 0$$

3. In a continuous distribution, whose p.d.f. is given by $f(x) = \frac{3}{4} x(2-x)$, $0 \leq x \leq 2$.

Show that the arithmetic mean of the distribution is 1 and variance is $\frac{1}{5}$

6.6 Moment Generating Function

Definition: The moment generating function (m.g.f.) for any random variable X about the origin is defined by $M_x(t) = E(e^{tx}) = \begin{cases} \int e^{tx} f(x) dx & \text{if } X \text{ is a continuous random variable with p.d.f. } f(x) \\ \sum_x e^{tx} p(x) & \text{if } X \text{ is a discrete random variable where the integration or summation is taken over the entire range of } X \text{ and } t \text{ is a real parameter.} \end{cases}$

$$M_x(t) = E(e^{tx}) = E\left[1 + \frac{tX}{1!} + \frac{t^2 X^2}{2!} + \dots + \frac{t^r X^r}{r!} + \dots\right]$$

$$= 1 + t E(X) + \frac{t^2}{2!} E(X^2) + \dots + \frac{t^r}{r!} E(X^r) + \dots$$

$$M_x(t) = E(e^{tx}) = 1 + t\mu_1^1 + \frac{t^2}{2!} \mu_2^1 + \dots + \frac{t^r}{r!} \mu_r^1 + \dots$$

where $\mu_r^1 = r^{\text{th}}$ moment about the origin.

$$= E(X^r) = \int x^r f(x) dx \text{ or}$$

$$= \sum x^r p(x) \text{ depending upon } X \text{ is continuous or discrete.}$$

Thus the coefficient of $\frac{t^r}{r!}$ in $M_x(t) = E(e^{tx})$ gives μ_r^1 . Thus $M_x(t)$ generates moments about the origin and hence the name moment generating function.

Note: 1. Differentiating (1) with respect to t , r times and setting $t = 0$ we get

$$\mu_r^1 = \frac{d^r}{dt^r} [M_x(t)]_{t=0}$$

2. In general, the m.g.f. of X about a point $X = a$ is defined as $M_x(t) \text{ (about } X=a) = E[e^{t(x-a)}]$

$$= E\left[1 + \frac{t(x-a)}{1!} + \frac{t^2 (x-a)^2}{2!} + \dots + \frac{t^r (x-a)^r}{r!} + \dots\right]$$

$$= 1 + \frac{t}{1!} \mu_1^1 + \frac{t^2}{2!} \mu_2^1 + \dots + \frac{t^r}{r!} \mu_r^1 + \dots$$

where $\mu_r^1 = E[(x-a)^r]$, r^{th} moment about the point $x = a$.

3. Taking a as mean then

$$M_x(t) \text{ (about } x = \bar{x}) = 1 + \frac{t}{1!} \mu_1 + \frac{t^2}{2!} \mu_2 + \dots + \frac{t^r}{r!} \mu_r + \dots$$

where $\mu_r = E[(x-\bar{x})^r] = r^{\text{th}}$ central moment.

Limitations of m.g.f.

1. A random variable X may have no moments although its m.g.f. exists.
2. A random variable X can have m.g.f. and some or all moments; yet the m.g.f. does not generate the moments.
3. A random variable X can have all or some moments; but m.g.f. does not exist except perhaps at one point.

Properties of moment generating function

$$1. M_{x+a}^{(t)} = e^{-at} M_x^{(t)}$$

Proof:

$$\begin{aligned} M_{x+a}^{(t)} &= E[e^{t(x+a)}] = E[e^{-at} e^{tx}] \\ &= e^{-at} E[e^{tx}] = e^{-at} M_x^{(t)} \end{aligned}$$

$$2. \text{ If } M_x(t) \text{ is the m.g.f. of } X \text{ and } Y = \alpha x + \beta \text{ then m.g.f. of } Y \text{ is } M_y(t) = e^{\beta t} M_x(\alpha t)$$

Proof: $M_y(t) = E(e^{ty}) = E[e^{(\alpha x + \beta)t}]$

$$\begin{aligned} &= E[e^{\alpha x t} e^{\beta t}] \\ &= e^{\beta t} E[e^{\alpha t x}] = e^{\beta t} M_x(\alpha t) \end{aligned}$$

$$3. \text{ If } X \text{ and } Y \text{ are two independent random variables and } Z = X + Y \text{ then } M_z(t) = M_x(t) M_y(t)$$

Proof: $M_z(t) = E(e^{zt}) = E[e^{t(x+y)}] = E(e^{tx} e^{ty})$

$$= E(e^{tx}) E(e^{ty}) \text{ (since } x \text{ and } Y \text{ are independent)}$$

$$= M_x(t) M_y(t).$$

If $X_1, X_2, \dots, X_n$ are n independent random variables then $M_{x_1} + M_{x_2} + \dots + M_{x_n}(t)$

$$= M_{x_1}(t) M_{x_2}(t) \dots M_{x_n}(t).$$

Example 1: X is a discrete random variable with probability function $p(x) = \frac{1}{k^x}$, $x = 1, 2, (k = \text{const})$

Find its (i) m.g.f. (ii) mean and (iii) variance.

$$\begin{aligned}
 \text{Solution: (i) } M_x(t) &= E(e^{tx}) = \sum_{x=1}^{\infty} e^{tx} \left(\frac{1}{k^x} \right) \\
 &= \sum_{x=1}^{\infty} \left(\frac{e^t}{k} \right)^x \\
 &= \frac{e^t}{k} \left[1 + \frac{e^t}{k} + \left(\frac{e^t}{k} \right)^2 + \dots \right] \\
 &= \frac{e^t}{k} \left[\frac{1}{1 - \frac{e^t}{k}} \right] = \frac{e^t}{k - e^t}
 \end{aligned}$$

$$(ii) \mu_1^1 = \frac{d}{dt} [M_x(t)]_{t=0} = \frac{d}{dt} \left[\frac{e^t}{k - e^t} \right]_{t=0}$$

$$= \left[\frac{(k - e^t) e^t - e^t (-e^t)}{(k - e^t)^2} \right]_{t=0}$$

$$= \left[\frac{ke^t}{(k - e^t)^2} \right]_{t=0} = \frac{k}{(k - 1)^2}$$

$$\therefore \text{Mean} = \frac{k}{(k - 1)^2}$$

$$(iii) \text{Variance} = \mu_2^1 - (\mu_1^1)^2$$

$$\mu_2^1 = \frac{d^2}{dt^2} [M_x(t)]_{t=0}$$

$$= \left[\frac{(k - e^t)^2 ke^t - ke^t 2 (k - e^t) (-e^t)}{(k - e^t)^4} \right]_{t=0}$$

$$= \frac{k(k+1)}{(k-1)^3}$$

$$\therefore \text{Variance} = \mu_2^1 - (\mu_1^1)^2 = \frac{k(k+1)}{(k-1)^3} - \frac{k^2}{(k-1)^4}$$

$$= \frac{k^3 - k^2 - k}{(k-1)^4}$$

Example 2: Let X have the probability mass function

$$f(k) = \begin{cases} \frac{6}{\pi^2} \frac{1}{k^2} & k = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

Find the moment generating function.

Solution:
$$M_x(t) = \sum_{k=1}^{\infty} e^{tk} \frac{6}{\pi^2} \frac{1}{k^2}$$

$$= \frac{6}{\pi^2} \sum_{k=1}^{\infty} e^{tk} \frac{1}{k^2}$$

$M_x(t)$ is infinity for any $t > 0$

$\therefore M_x(t)$ does not exist

Example 3: Find the m.g.f. and the r^{th} moment for the distribution whose p.d.f. is $f(x) = ke^{-x}$, $0 \leq x \leq \alpha$.

Solution:
$$\int_0^{\infty} ke^{-x} dx = 1 \quad [-ke^{-x}]_0^{\infty} = 1$$

$$\therefore k=1$$

$$M_x(t) = E(e^{tx}) = \int_0^{\infty} e^{tx} e^{-x} dx.$$

$$= \int_0^{\infty} e^{(t-1)x} dx$$

$$= \left[\frac{e^{(t-1)x}}{t-1} \right] = \frac{1}{1-t} \text{ if } t < 1$$

$$= (1-t)^{-1} = 1 + t + t^2 + t^3 + \dots + t^r + \dots$$

$\mu_r^1 = \text{Coefficient of } \frac{t^r}{r!} \text{ is } r!$

$$\therefore \mu_1^1 = 1; \mu_2^1 = 2.$$

$$\therefore \text{Mean} = 1$$

$$\text{Variance} = \mu_2 - (\mu_1)^2$$

$$= 2 - (1)^2 = 1$$

Exercise:

1. A random variable X has the p.d.f.

$$f(x) = \begin{cases} \frac{1}{2x} & \text{for } x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

Find its m.g.f., mean and variance.

2. A random variable X has p.d.f.

$$f(x) = \begin{cases} 2e^{-2x} & , x \geq 0 \\ 0 & , x < 0 \end{cases} \text{ obtain the m.g.f.}$$

and the first four moments about the origin.

Unit 7

Special Distribution

In this unit we discuss some important distributions of random variables which are frequently used in Statistics. They are:

1. Binomial distribution
2. Poission distribution
3. Normal distribution
4. Gamma distribution
5. Chisquare distribution
6. Student - t distribution
7. Snedecor's F - distribution
8. Fischer's Z - distribution

Among these, the first two are of discrete type and the other six are of continuous type.

7.1 - Binomial distribution

The binomial distribution also known as 'Bernoulli distribution' is associated with the name of a Swiss Mathematician James Bernoulli. Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives ie., success or failure.

This distribution has been used to describe a wide variety of processes in business and the social sciences as well as other areas. The mathematical model for a Bernoulli process is developed under a very specific set of assumption involving the concept of a series of experimental trials. These assumptions are:

1. An experiment is performed under the same conditions for a fixed number of trials say, n .
2. In each trial, there are only two possible outcomes of the experiment, success or failure.
3. The probability of a success denoted by p remains constant from trial to trial. The probability of a failure denoted by q is equal to $(1-p)$. If the probability of success is not the same in each trial, we will not have binomial distribution.
4. The trials are statistically independent, ie., the outcomes of any trial or sequence of trials do not affect the outcomes of subsequent trials.

We now find $P(X=x)$ where $0 \leq x \leq n$. $X=x$ iff exactly x of the variables $X_1, X_2, \dots, X_n$ have value 1 and the remaining $n-x$ variables have value 0. There are nc_x ways in which exactly x variables take the value 1 and the probability of each of these nc_x ways is $p^x q^{n-x}$, since $X_1, X_2, \dots, X_n$ are independent random variables. Thus $P(X=x) = nc_x p^x q^{n-x}$.

Clearly, $\sum_{x=0}^n nc_x p^x q^{n-x} = (p+q)^n = 1$ so that this is the probability density function of a discrete random variable.

Definition: Let n be any positive integer and let $0 < p < 1$. Let $q=1-p$

$$\text{Define } p(x) = \begin{cases} nc_x p^x q^{n-x} & \text{if } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

A discrete random variable with the above p.d.f. is said to have binomial distribution and the p.d.f. itself is called a binomial distribution.

- Note: 1. The two independent constants n and p in the distribution are known as the parameters of the distribution.
2. The probability distribution function of a binomial distribution is obtained by considering n independent trials of a random experiment whose outcome is success or failure. If this experiment is repeated N times then the frequency function of the binomial distribution is given by

$$f(x) = N p(x) = N nc_x p^x q^{n-x}; x=0, 1, 2, \dots, n$$

and the expected frequencies of 0, 1, 2, ..., n successes are the successive terms of the binomial expansion $N(q+p)^n$ where $p+q=1$.

x	Probability	Expected frequency
0	q^n	Nq^n
1	$nc_1 pq^{n-1}$	$N nc_1 pq^{n-1}$
2	$nc_2 p^2 q^{n-2}$	$N nc_2 p^2 q^{n-2}$
..		
n	p^n	Np^n

Theorem 1: M.g.f. of a binomial distribution about the origin is $(q+pe^t)^n$

Proof: $M_x(t) = E(e^{xt}) = \sum_x e^{xt} p(x)$

$$= \sum_{x=0}^n e^{tx} nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^n nC_x (pe^t)^x \cdot q^{n-x}$$

$$= (q+pe^t)^n$$

Moments of binomial distribution:

We know that for any random variable X the m.g.f. is $M_x(t) = 1 + \mu_1 t + \frac{\mu_2^1}{2!} t^2 + \frac{\mu_3^1}{3!} t^3 + \dots$

$$+ \frac{\mu_r^1}{r!} t^r + \dots \quad (1)$$

For binomial variate $M_x(t) = (q+pe^t)^n$

$$= [q + p (1 + \frac{t}{1!} + \frac{t^2}{2!} + \dots)]^n$$

$$= [q + p + p (\frac{t}{1!} + \frac{t^2}{2!} + \dots)]^n$$

$$= [1 + p (\frac{t}{1!} + \frac{t^2}{2!} + \dots) + \frac{n(n-1)}{2!} p^2 (\frac{t}{1!} + \frac{t^2}{2!} + \dots)^2$$

$$+ \frac{n(n-1)(n-2)}{3!} p^3 (\frac{t}{1!} + \frac{t^2}{2!} + \dots)^3 + \dots]$$

$$= 1 + \frac{(np)t}{1!} + [np + n(n-1)p^2] \frac{t^2}{2!} + [np + 3n(n-1)p^2 + n(n-1)(n-2)p^3] \frac{t^3}{3!} + \dots \quad (2)$$

Comparing (1) and (2), we get

$$\mu_1^1 = np = \mu_1 \text{ (mean)}$$

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

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

Note: Since $\mu_1 = \mu = \text{mean}$, we have mean of the binomial distribution is np.

m.g.f. about mean np of a binomial distribution is $(qe^{-pt} + pe^{qt})^n$

Proof: $M_x(t) = M_{x=np}(t) = E[e^{t(x-np)}] = e^{-npt} M_x(t)$

$$= e^{-npt} (q+pe^t)^n$$

$$= [qe^{-pt} + pe^{te^{-pt}}]^n = [qe^{-pt} + p(e)^{(1-p)t}]^n$$

$$= (qe^{-pt} + pe^{qt})^n$$

Using m.g.f. about the mean of the binomial distribution we can find central moments $\mu_2, \mu_3, \mu_4, \dots$

We know that for any random variable X the m.g.f about

$$X = m \text{ is } M_x =_m(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \dots \quad (1)$$

For binomial variable x we have $M_{x=np}(t) = (qe^{-pt} + pe^{qt})^n$

$$= [q(1 - \frac{pt}{2!} + \frac{p^2 t^2}{2!} - \frac{p^3 t^3}{3!} + \frac{p^4 t^4}{4!} - \dots) + p(1 + \frac{tq}{1!} + \frac{t^2 q^2}{2!} + \frac{t^3 q^3}{3!} + \frac{t^4 q^4}{4!} + \dots)]^n$$

$$= [(q+p) + (p+q)pq \frac{t^2}{2!} + pq(q^2 - p^2) \frac{t^3}{3!} + pq(p^3 + q^3) \frac{t^4}{4!} + \dots]^n$$

$$= [1 + pq \frac{t^2}{2!} + pq(p-q) \frac{t^3}{3!} + pq(p^2 - pq + q^2) \frac{t^4}{4!} + \dots]^n$$

$$= 1 + n \{pq \frac{t^2}{2!} + pq(p-q) \frac{t^3}{3!} + pq(p^2 - pq + q^2) \frac{t^4}{4!} + \dots\}$$

$$+ \frac{n(n-1)}{2!} \{pq \frac{t^2}{2!} + \dots\}^2 + \dots \text{ (expanding in binomial series)}$$

$$= 1 + npq \frac{t^2}{2!} + npq(q-p) \frac{t^3}{3!} + [npq(1-3pq) \frac{t^4}{4!} + \frac{n(n-1)}{2!} p^2 q^2 \frac{t^4}{4!}] + \dots$$

$$= 1 + npq \frac{t^2}{2!} + npq(q-p) \frac{t^3}{3!} + [npq(1-3pq) + 3n(n-1)p^2 q^2] \frac{t^4}{4!} + \dots \quad (2)$$

From (1) and (2) we get $\mu_2 = npq$

$$\mu_3 = npq(q-p)$$

$$\mu_4 = npq(1-6pq) + 3n^2 p^2 q^2$$

Theorem

Addition property of binomial distribution

If X_1 is a binomial variate with parameters n_1 and p and X_2 is a binomial variate with parameters n_2 and p and X_1 and X_2 are independent random variables then $X_1 + X_2$ is a binomial variate with parameters $n_1 + n_2$ and p .

Proof: Let us consider m.g.f. of X_1 and X_2 about origin.

$$\therefore M_{X_1}(t) = (q+pe^t)^{n_1} \text{ and } M_{X_2}(t) = (q+pe^t)^{n_2}$$

$$M_{X_1 + X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \quad (\because X_1 \text{ and } X_2 \text{ are independent})$$

$$= (q+pe^t)^{n_1} (q+pe^t)^{n_2} = (q+pe^t)^{n_1 + n_2}$$

= m.g.f. of the binomial $X_1 + X_2$ with parameters $n_1 + n_2$ and p .

Hence $X_1 + X_2$ is a binomial variable with parameters $n_1 + n_2$ and p .

When we write binomial expansion, the following general relationships should be noted.

1. The number of terms in a binomial expansion is always $n+1$.
2. The exponents of p and q , for any single term, when added together always sum to n .
3. The exponents of q are $n, (n-1), (n-2), \dots, 1, 0$ respectively and the exponents of p are $0, 1, 2, \dots, (n-1), n$ respectively.
4. The coefficients for the $n+1$ terms of the distribution are always symmetrical ascending to the middle of the series and then descending, when n is odd number, $n+1$, is even and the coefficients of the two central terms are identical.

Mode of binomial distribution

Let X be a binomial variate with parameters n and p . Then $p(x) = {}^n C_x p^x q^{n-x}$. Let x be the mode of the binomial distribution.

Since mode is the value of x for which $p(x)$ is maximum we have $p(x) > p(x+1)$ and $p(x) > p(x-1)$.

$$\text{Now } p(x) > p(x+1) \Rightarrow \frac{p(x)}{p(x+1)} \geq 1.$$

$$\Rightarrow \left(\frac{x+1}{n-x} \right) \left(\frac{q}{p} \right) \geq 1 \quad [\because p(x+1) = {}^n C_{x+1} p^{x+1} q^{n-x-1}]$$

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

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

$$= \left(\frac{n-x}{x+1} \right) \left(\frac{p}{q} \right) \frac{n! p^x q^{n-x}}{x! (n-x)!}$$

$$= \left(\frac{n-x}{x+1} \right) \left(\frac{p}{q} \right) p(x)$$

$$\Rightarrow (x+1) q \geq (n-x) p$$

$$\Rightarrow qx + q \geq np - px$$

$$\Rightarrow px + qx \geq np - q$$

$$\Rightarrow (p+q)x \geq np - (1-p)$$

$$\Rightarrow x \geq np - 1 + p$$

$$\Rightarrow x \geq (n+1)p - 1.$$

Taking $p(x) \geq p(x-1)$ we can get $x \leq (n+1)p$.

$$\therefore (n+1)p - 1 \leq x \leq (n+1)p.$$

Case(i): If $(n+1)p$ is not an integer clearly mode is the integral part of $(n+1)p$ and the distribution is unimodel.

Case(ii): If $(n+1)p$ is an integer both $(n+1)p$ and $(n+1)p - 1$ will represent the mode and the distribution is bimodel.

Importance of the Binomial distribution

1. The outcome or results of each trial in the process are characterised as one of two types of possible outcomes. In other words they are attributes.
2. The possibility of outcome of any trial does not change and is independent of the results of previous trials.

Example 1: A coin is tossed six times. What is the probability of obtaining four or more heads?

Solution: When a coin is tossed the probabilities of head and tail in case of an unbiased coin are equal ie., $p = q = \frac{1}{2}$

The various possibilities for all the events are the terms of the expansion $(q+p)^6$

$$(q+p)^6 = q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 + 15q^2p^4 + 6qp^5 + p^6$$

$\therefore$ The probability of obtaining 4 heads is

$$15q^2p^4 = 15 \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^4 = 0.234$$

The probability of obtaining 5 heads is

$$6qp^5 = 6\left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^5 = 0.094$$

The probability of obtaining 6 heads is

$$(q)^6 = \left(\frac{1}{2}\right)^6 = 0.016$$

$\therefore$ The probability of obtaining 4 or more heads is

$$0.234 + 0.094 + 0.016 = 0.344$$

Example 2: Six dice are thrown 729 times. How many times do you expect at least 3 dice to show a five or six.

Solution: Here $n = 6$; $N = 729$.

p = probability of getting 5 or 6 with one die

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

$$\text{Hence } q = \frac{2}{3}$$

The expected frequencies of 0, 1, 2, ..., 6 successes are the successive terms of

$$729 \left(\frac{1}{3} + \frac{2}{3} \right)^6$$

$\therefore$ Expected number of times atleast 3 dice showing five or six.

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

$$= \frac{729}{3^6} [20 \times 2^3 + 15 \times 2^2 + 6 \times 2 + 1]$$

$$= \frac{729}{3^6} \times 233 = 233.$$

Fitting a Binomial Distribution

When a binomial distribution is to be fitted to observe data, the following procedure is adopted.

1. Determine the values of p and q . If one of these values is known the other can be found out by the simple relationship $p = (1-q)$ and $q = (1-p)$. When p and q are equal the distribution is symmetrical. If p and q are unequal, the distribution is skew. If p is less than $\frac{1}{2}$, the distribution is positively skewed and when p is more than $\frac{1}{2}$ the distribution is negatively skewed.
2. Expand the binomial $(q+p)^n$. The power n is equal to one less than the number of terms in the expanded binomial. Thus when two coins are tossed there will be three terms in the binomial. When four coins are tossed, there will be five terms and so on.
3. Multiply each term of the expanded binomial by N (the total frequency), in order to obtain the expected frequency in each category.

Example 1: Eight coins are tossed at 256 times. Number of heads observed at each throw is recorded and the results are given below. Find the expected frequencies. What are the theoretical values of mean and standard deviation?

No. of heads at a throw	Frequency
0	2
1	6
2	30
3	52
4	67
5	56
6	32
7	10
8	1

Solution: The chance of getting a head in a single throw of one coin is $\frac{1}{2}$.

$$\therefore p = \frac{1}{2}, \quad q = \frac{1}{2}, \quad n = 8, \quad N = 256.$$

By expanding $256 \left(\frac{1}{2} + \frac{1}{2} \right)^8$ we shall get the expected frequencies of 1, 2,, 8 heads.

No. of heads (X)	Frequency - $N \times {}^n C_r \cdot q^{n-r} p^r$
0	$256 \left(\frac{1}{2} \right)^8 = 1$
1	$256 \times 8 {}^8 C_1 \left(\frac{1}{2} \right)^1 \left(\frac{1}{2} \right)^7 = 8$
2	$256 \times 8 {}^8 C_2 \left(\frac{1}{2} \right)^2 \left(\frac{1}{2} \right)^6 = 28$
3	$256 \times 8 {}^8 C_3 \left(\frac{1}{2} \right)^3 \left(\frac{1}{2} \right)^5 = 56$
4	$256 \times 8 {}^8 C_4 \left(\frac{1}{2} \right)^4 \left(\frac{1}{2} \right)^4 = 70$
5	$256 \times 8 {}^8 C_5 \left(\frac{1}{2} \right)^5 \left(\frac{1}{2} \right)^3 = 56$
6	$256 \times 8 {}^8 C_6 \left(\frac{1}{2} \right)^6 \left(\frac{1}{2} \right)^2 = 28$
7	$256 \times 8 {}^8 C_7 \left(\frac{1}{2} \right)^7 \left(\frac{1}{2} \right)^1 = 8$
8	$256 \left(\frac{1}{2} \right)^8 = 1$
Total	= 256

$$\text{mean} = np = 8 \times \frac{1}{2} = 4.$$

$$\text{S.D.} = \sqrt{npq} = \sqrt{8 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{2} = 1.414$$

Example 2: The following data show the number of seeds germinating out of 10 on damp filter for 80 set of seeds. Fit a binomial distribution to this data.

x:	0	1	2	3	4	5	6	7	8	9	10
f:	6	20	28	12	8	6	0	0	0	0	0

Solution:

Fitting Binomial Distribution

x	f	fx
0	6	0
1	20	20
2	28	56
3	12	36
4	8	32
5	6	30
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0

$$N=80$$

$$\Sigma fx = 174$$

$$\bar{x} = \frac{174}{80} = 2.175$$

$$\text{But mean} = np = \frac{174}{80}$$

$$n = 10 \quad \therefore 10p = \frac{174}{80}$$

$$p = \frac{174}{800} = 0.2175$$

$$\therefore q = 1 - p = 0.7825$$

Hence the binomial distribution to be fitted to the data is $80(0.7825 + 0.2175)^{10}$

The theoretical frequencies are the successive terms in the expansion of $80(0.7825 + 0.2175)^{10}$ and are tabulated below.

X	Theoretical frequencies NX nc, $q^{n-r} p^r$
0	$80 \times (0.7825)^{10} = 6.9$
1	$80 \times 10 (0.7825)^9 (0.2175)^1 = 19.1$
2	$80 \times 45 (0.7825)^8 (0.2175)^2 = 24.0$
3	$80 \times 120 (0.7825)^7 (0.2175)^3 = 17.8$
4	$80 \times 210 (0.7825)^6 (0.2175)^4 = 8.6$
5	$80 \times 252 (0.7825)^5 (0.2175)^5 = 2.9$
6	$80 \times 210 (0.7825)^4 (0.2175)^6 = 0.7$
7	$80 \times 120 (0.7825)^3 (0.2175)^7 = 0.1$
8	$80 \times 45 (0.7825)^2 (0.2175)^8 = 0.0$
9	$80 \times 10 (0.7825)^1 (0.2175)^9 = 0.0$
10	$80 \times (0.2175)^{10} = 0.0$
Total = 80.1	

Example 3: Find β and r coefficients for the binomial distribution and discuss the results with special reference to skewness and kurtosis.

Solution: For a binomial distribution,

$$\mu_2 = npq.$$

$$\mu_3 = npq(q-p)$$

$$\text{and } \mu_4 = npq(1-6pq) + 3n^2p^2q^2$$

$$\text{We know } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{n^2p^2q^2(q-p)^2}{(npq)^3} = \frac{(q-p)^2}{npq}$$

$$\begin{aligned}\beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{npq(1-6pq) + 3n^2p^2q^2}{(npq)^2} \\ &= 3 + \frac{1-6pq}{npq}\end{aligned}$$

$$r_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} = \frac{1-2p}{\sqrt{npq}}$$

$$r_2 = \beta_2 - 3 = \frac{1-6pq}{npq}$$

We know r_1 is a measure for skewness.

$$r_1 = \frac{1-2p}{\sqrt{npq}} \Rightarrow r_1 = 0 \text{ when } p = \frac{1}{2}$$

$$r_1 < 0 \text{ when } p > \frac{1}{2} \text{ and } r_1 > 0 \text{ when } p < \frac{1}{2}.$$

$\Rightarrow$ Skewness is negative, zero or positive according as $p > \frac{1}{2}$ or $p = \frac{1}{2}$

$$\text{or } p < \frac{1}{2}.$$

We know r_2 is a measure of kurtosis.

$$r_2 = \frac{1-6pq}{npq} \Rightarrow r_2 < 0 \text{ when } pq > \frac{1}{6}$$

$$r_2 = 0 \text{ when } pq = \frac{1}{6} \text{ and } r_2 > 0 \text{ when } pq < \frac{1}{6}$$

$\Rightarrow$ The curve is platykurtic or mesokurtic or leptokurtic according as $pq < \frac{1}{6}$ or $pq = \frac{1}{6}$

$$\text{or } pq > \frac{1}{6}.$$

Exercise

1. For a binomial distribution the mean is 6 and the S.D. is $\sqrt{2}$. Write down all the terms of the distribution.
2. The mean and variance of a binomial variate X with parameters n and p are 16 and 8 respectively. Find (i) $P(X=0)$ (ii) $P(X=1)$ (iii) $P(X \geq 2)$.
3. 10 unbiased coins are thrown simultaneously. Find the probability of getting at least seven heads.
4. Obtain the M.g.f. of the binomial distribution with $n=7$ and $p=0.6$. Find the first three moments of the distribution.

7.2 Poisson Distribution

Poisson distribution is a discrete probability distribution and is very widely used in statistical work. It was developed by a French mathematician Simeon Denis Poisson (1781-1840) in 1837. We know that binomial distribution is determined by two parameters p and n . If the number of trials is indefinitely large and the probability p of success is infinitesimally small such that $np = \lambda$ where λ is a constant then the limiting case of the binomial distribution when $n \rightarrow \infty$ and $p \rightarrow 0$ becomes a distribution known as **Poisson distribution**.

Definition: A discrete random variable X is said to follow a Poisson distribution if it assumes only non negative integer values and its probability density function is given by

$$p(x) = P(X=x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & \text{if } x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

where λ is a parameter of the distribution.

The Poisson distribution is attributable in the case of the rare events.

The following are some instances where Poisson distribution may be employed.

- (i) Number of deaths due to rare disease such as snake bite, cancer.
- (ii) The number of defective articles in a packing manufactured by a good company.
- (iii) The number of telephone calls received in some unit of time.
- (iv) The number of cars passing a place per minute during peak hours per day.
- (v) The number of children born blind per year in a city.

Role of the Poisson distribution

The Poisson distribution is used in practice in a wide variety of problems where there are infrequently occurring events with respect to time area, volume or similar units. Some practical situations in which Poisson distribution can be used are given below:

- 1) It is used in quality control statistics to count the number of defects of an item.
- 2) In biology to count the number of bacteria.
- 3) In physics to count the number of particles emitted from a radioactive substance.

4) In insurance problems to count the number of casualties.

5) To model the distribution of the number of persons joining a queue to receive a service or purchase of a product.

Theorem 1: Mean of the Poisson distribution is λ .

Proof: Let X be a random Poisson variate

$$\begin{aligned}\text{Mean} = E(X) &= \sum_{x=0}^{\infty} x P(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} \\ \mu_1' &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \dots \right] \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda.\end{aligned}$$

Hence mean of the Poisson distribution is λ .

Theorem 2: The standard deviation of the Poisson distribution is $\sqrt{\lambda}$.

Proof: Consider $\mu_2' = E(X^2) = \sum_{x=0}^{\infty} x^2 p(x)$

$$\begin{aligned}&= \sum_{x=0}^{\infty} \frac{x^2 e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{[x(x-1) + x] \lambda^x}{x!} \\ &= e^{-\lambda} \left[\sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} + \sum_{x=0}^{\infty} \frac{x \lambda^x e^{-\lambda}}{x!} \right] \\ &= e^{-\lambda} \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + E(X) \\ &= e^{-\lambda} \lambda^2 e^{\lambda} + \lambda\end{aligned}$$

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

$$\therefore \mu_2 = \mu_2' - (\mu_1')^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda.$$

$$\text{Variance} = \lambda.$$

$$\therefore \text{S.D.} = \sqrt{\lambda}.$$

Mode of the Poisson distribution:

We know mode is the value of x for which the probability $p(x)$ is maximum.

$$\therefore p(x) \geq p(x+1) \text{ and } p(x) \geq p(x-1)$$

Now $p(x) \geq p(x+1)$

$$\begin{aligned} \Rightarrow \frac{p(x)}{p(x+1)} &\geq 1 \Rightarrow \frac{x+1}{\lambda} \geq 1. \quad [\because p(x+1) = \frac{e^{-\lambda} \lambda^{x+1}}{(x+1)!}] \\ &= \frac{e^{-\lambda} \lambda^x \lambda}{(x+1) x!} \\ \Rightarrow x+1 &\geq \lambda \quad \quad \quad = \frac{\lambda}{(x+1)} \left(\frac{e^{-\lambda} \lambda^x}{x!} \right) \\ \Rightarrow x &\geq \lambda - 1. \quad \quad \quad = \left(\frac{\lambda}{x+1} \right) p(x) \end{aligned}$$

Similarly, $p(x) \geq p(x-1) \rightarrow x \leq \lambda$

$$\therefore \lambda - 1 \leq x \leq \lambda.$$

If λ is an integer $\lambda - 1$ is also an integer then mode has both the values $\lambda - 1$ and λ i.e., bimodal.

If λ is not an integer, mode is the integral part of λ which is the unique mode.

M.g.f. of the Poisson distribution about $x = 0$ is $e^\lambda (e^t - 1)$.

Proof: $M_x(t) = E(e^{tx}) = \sum_x e^{tx} p(x)$

$$= \sum_{x=0}^{\infty} \frac{e^{tx} e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

Corollary 1: M.g.f. of Poisson distribution about mean is $e^{\lambda(e^t - 1)}$

$$M_{x-\lambda}(t) = e^{-\lambda t} M_x(t) = e^{-\lambda t} e^{\lambda(e^t - 1)}$$

$$= e^{\lambda(e^t - 1 - t)}$$

Corollary 2: First four moments of the Poisson distribution about mean can be found out from m.g.f. about mean.

We know $M_{x=\text{mean}}(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \dots$

$$M_{x=\lambda}(t) = e^{\lambda(e^t - 1 - t)} = e^{\lambda \left(\frac{t^2}{2!} + \frac{t^3}{3!} + \dots \right)} \dots\dots\dots(1)$$

$$= 1 + \lambda \left(\frac{t^2}{2!} + \frac{t^3}{3!} + \dots \right) + \frac{\lambda^2}{2!} \left(\frac{t^2}{2!} + \frac{t^3}{3!} + \dots \right)^2$$

$$= 1 + \frac{\lambda t^2}{2!} + \frac{\lambda t^3}{3!} + \left(\frac{\lambda}{4!} + \frac{\lambda^2}{8} \right) \frac{t^4}{4!} + \dots$$

$$= 1 + \frac{\lambda t^2}{2!} + \frac{\lambda t^3}{3!} + (\lambda + 3\lambda^2) \frac{\lambda t^4}{4!} + \dots\dots\dots(2)$$

From (1) and (2) we get $\mu_2 = \lambda$

$$\mu_3 = \lambda$$

$$\text{and } \mu_4 = 3\lambda^2 + \lambda$$

β and r coefficients for Poisson distribution

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

$$\beta^2 = \frac{\mu_4}{\mu_2^3} = \frac{3\lambda^2 + \lambda}{\lambda^3} = 3 + \frac{1}{\lambda}$$

$$r_1 = \sqrt{\beta_1} = \sqrt{\frac{1}{\lambda}}$$

$$r_2 = \beta_2^{-3} = \frac{1}{\lambda}$$

Theorem3: Additive property of independent Poisson variates

If X_1 and X_2 are two independent Poisson variates with parameters λ_1 and λ_2 then the sum $X_1 + X_2$ is also a Poisson variate with parameter $\lambda_1 + \lambda_2$.

Proof: Given X_1 and X_2 are Poisson variates with parameters λ_1 and λ_2 .

$$\therefore M_{x_1}(t) = e^{\lambda_1(e^t - 1)} \quad \text{and } M_{x_2}(t) = e^{\lambda_2(e^t - 1)}$$

Since X_1 and X_2 are independent random variables

we have $M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t)$

$$= e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)}$$

$$= e^{\lambda_1(e^t-1) + \lambda_2(e^t-1)}$$

$$= e^{(\lambda_1+\lambda_2)(e^t-1)} \text{ which is}$$

m.g.f. of a Poisson variate with parameter $\lambda_1 + \lambda_2$.

$\therefore X_1 + X_2$ is a Poisson variate with parameter $\lambda_1 + \lambda_2$.

Theorem 4: The limiting form of the binomial distribution as $n \rightarrow \infty$ and $p \rightarrow 0$ in such a way that $np = \lambda$ where λ is a fixed positive number is the Poisson distribution.

Proof: Let X be a Binomial variate.

$$\therefore p(x) = {}^nC_x p^x q^{n-x}$$

$$= {}^nC_x p^x (1-p)^{n-x} \quad (\because p+q=1).$$

$$= {}^nC_x p^x \frac{(1-p)^n}{(1-p)^x}$$

$$= \frac{n(n-1)(n-2)\dots(n-x+1) \left(\frac{\lambda}{n}\right)^x}{x! \left(1 - \frac{\lambda}{n}\right)^x}$$

$$(\because np = \lambda \Rightarrow p = \frac{\lambda}{n})$$

$$= \frac{n^x \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots [1 - (x-1)/n] \lambda^x (1 - \lambda/n)^n}{x! \left(1 - \frac{\lambda}{n}\right)^x n^x}$$

For fixed x as $n \rightarrow \infty$ we know that

$$1 - \frac{1}{n}, 1 - \frac{2}{n}, \dots, [1 - (x-1)/n] \text{ all tend to } 1.$$

$$\left(1 - \frac{\lambda}{n}\right)^x \rightarrow 1 \text{ and } \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}$$

$$\therefore \text{Lt}_{n \rightarrow \infty} p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ which is the p.d.f. of the Poisson variate } x. \text{ Hence the theorem.}$$

Fitting of Poisson distribution to the observed data

The process of fitting a Poisson distribution is very simple. We have just to obtain the value of λ i.e., the average occurrence, and calculate the frequency of 0 success. The other frequencies can be very easily calculated as follows:

$$N(P_0) = Ne^{-\lambda}$$

$$N(P_1) = N(P_0) \times \frac{\lambda}{1}$$

$$N(P_2) = N(P_1) \times \frac{\lambda}{1}$$

$$N(P_3) = N(P_2) \times \frac{\lambda}{1} \text{ etc.}$$

Example 1: The following mistakes per page were observed in a book:

No. of mistakes per page : 0 1 2 3 4

No. of times the mistake occurred : 211 90 19 5 0

Fit a Poisson distribution to fit data.

Solution: Fitting Poisson distribution

X	f	fx
0	211	0
1	90	90
2	19	38
3	5	15
4	0	0
N=325		$\Sigma fx = 143$

$$\bar{X} = \frac{\Sigma fx}{N} = \frac{143}{325} = 0.44$$

Mean of the distribution or $\lambda = 0.44$

$$P_0 = e^{-\lambda} = (2.7183)^{-0.44}$$

$$= 0.6443.$$

$$N(P_0) = 0.6443 \times 325 = 209.40$$

$$N(P_1) = N(P_0) \times \frac{\lambda}{1} = 209.4 \times 0.44 = 92.14$$

$$N(P_2) = N(P_1) \times \frac{\lambda}{2} = 92.14 \times \frac{0.44}{2} = 20.27$$

$$N(P_3) = N(P_2) \times \frac{\lambda}{3} = 20.27 \times \frac{0.44}{3} = 2.97$$

$$N(P_4) = N(P_3) \times \frac{\lambda}{4} = 2.97 \times \frac{0.44}{4} = 0.33$$

The expected frequencies of Poisson distribution are:

X	:	0	1	2	3	4
f	:	209.4	92.14	20.27	2.97	0.33

Example 2: The following table gives the number of days in a 50 day period during which automobile accidents occurred in a certain part of a city. Fit a Poisson distribution to the data.

No. of accidents : 0 1 2 3 4

No. of days : 19 18 8 4 1

Solution: Fitting of Poisson distribution

X	f	fx
0	19	0
1	18	18
2	8	16
3	4	12
4	1	4
N=50		$\Sigma fx = 50$

$$\lambda = \frac{\Sigma fx}{N} = \frac{50}{50} = 1.$$

$$P_0 = e^{-\lambda} = (2.7183)^{-1} = 0.36788$$

$$N(P_0) = 50 \times 0.36788 = 18.394 = 18.4$$

$$N(P_1) = N(P_0) \times \lambda = 18.394 \times 1 = 18.4.$$

$$N(P_2) = N(P_1) \times \frac{\lambda}{2} = \frac{18.394}{2} = 9.19 = 9.2.$$

$$N(P_3) = N(P_2) \times \frac{\lambda}{2} = \frac{9.19}{3} = 3.066 = 3.1.$$

$$N(P_4) = N(P_3) \times \frac{\lambda}{2} = \frac{3.066}{4} = 0.766 = 0.88.$$

The expected frequencies of Poisson distribution are:

X	:	0	1	2	3	4
Y	:	18.4	18.4	9.2	3.1	0.8

Exercise

1. The number of defects per unit in a sample of 330 units of manufactured product was found as follows:

No. of defects : 0 1 2 3 4

No. of Units : 214 92 20 3 1

Fit a Poisson distribution to the data. ($e^{-0.439} = 0.6447$)

2. If X is a Poisson variate such that $P(X=1) = P(X=2)$ find $P(X=4)$.
3. If X is a Poisson variate with parameter 1 find $P(3 < X < 5)$.
4. The mean of a Poisson distribution is 2.25.

Find (i) σ (ii) μ_3 , (iii) β_1 , (iv) β_2 .

5. Fit a Poisson distribution to the following data.

X	0	1	2	3	4
f	109	665	22	3	1

7.3 Normal distribution

The normal distribution also called the normal probability distribution happens to be most useful theoretical distribution for continuous variables. Many statistical data concerning business and economic problems are displayed in the form of normal distribution. Infact normal distribution is the cornerstone of modern statistics.

The normal distribution is an approximation to binomial distribution. Whether or not p is equal to q , the binomial distribution tends to the form of the continuous curve and when n becomes large atleast for the material part of the range. The limiting frequency curve obtained as n becomes large is called the normal frequency curve or simply the normal curve.

Definition: A continuous random variable X is said to follow a normal distribution if its probability density function is given by

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

μ and σ are constants and $\sigma > 0$ and are called the parameters of the distribution.

Moment generating function about the origin of the Normal distribution is $e^{\mu t + (t^2 \sigma^2)/2}$

Proof: $M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2\sigma^2} dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(\sigma z + \mu)} e^{-z^2/2} \sigma dz \quad \left(\text{putting } z = \frac{x - \mu}{\sigma} \right)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{mt} e^{-(z^2)/2 + t\sigma z} dz$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z^2 - 2\sigma t z)/2} dz$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \{(z - \sigma t)^2 - \sigma^2 t^2\}} dz$$

$$= e^{\mu t + (t^2 \sigma^2)/2} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} du \quad \left(\text{putting } u = z - \sigma t \right)$$

$$= e^{\mu t + (t^2 \sigma^2)/2} \times \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = e^{\mu t + (t^2 \sigma^2)/2}$$

Mean of the normal distribution is μ

For any m.g.f. we know $\left[\frac{d^r}{dt^r} (M_x(t)) \right]_{t=0} = \mu_r^1$

$$\therefore \mu_1^1 = \left[\frac{d}{dt} (e^{\mu t + t^2 \sigma^2 / 2}) \right]_{t=0}$$

$$= [e^{\mu t + t^2 \sigma^2 / 2} (\mu + t \sigma^2)]_{t=0} = \mu.$$

$\therefore \mu$ is the mean of the normal distribution.

Variance of the Normal distribution is σ^2

$$\mu_2^1 = \left[\frac{d^2}{dt^2} (M_x(t)) \right]_{t=0}$$

$$= [e^{\mu t + t^2 \sigma^2 / 2} \sigma^2 + (\mu + t \sigma^2) e^{\mu t + t^2 \sigma^2 / 2} (\mu + t \sigma^2)]_{t=0}$$

$$= \sigma^2 + \mu^2$$

$$\therefore \mu_2 = \mu_2^1 - (\mu_1^1)^2 = (\sigma^2 + \mu^2) - \mu^2 = \sigma^2.$$

$\therefore$ Variance of the normal distribution is σ^2 .

Standard Normal Variate

Let X be a normal variate with parameters μ and σ^2 .

$$\text{Put } z = \frac{x - \mu}{\sigma}$$

$$\therefore E(z) = E\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma} [E(x) - E(\mu)]$$

$$= \frac{\mu - \mu}{\sigma} = 0.$$

$$\text{Variance of } z = \mu_2 = E(z^2) - [E(z)]^2$$

$$= E\left[\left(\frac{x - \mu}{\sigma}\right)^2\right] = \frac{1}{\sigma^2} E(x - \mu)^2 = 1.$$

$\therefore Z$ is a normal variate with mean 0 and variance 1. We say that Z is the standard normal variate. The p.d.f. of Z is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Mode of the normal distribution

Let X be a normal variate with mean μ and variance σ^2 . $\therefore f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

Mode is the value of x for which $f(x)$ has maximum.

$\therefore$ Mode is the value of x such that $f'(x)=0$ and $f''(x) < 0$.

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

Differentiating w.r.t. x , we get

$$\frac{f'(x)}{f(x)} = -\frac{1}{\sigma^2} (x-\mu)$$

$$\therefore f'(x) = -\frac{1}{\sigma^2} f(x) (x-\mu)$$

$$f''(x) = -\frac{1}{\sigma^2} [f'(x) (x-\mu) + f(x)]$$

$$= -\frac{f(x)}{\sigma^2} \left[1 - \frac{(x-\mu)^2}{\sigma^2} \right]$$

$$f'(x) = 0 \Rightarrow x = \mu \quad (\because \sigma > 0 \text{ and } f(x) \neq 0)$$

$$\text{At } x=\mu, f''(x) = -\frac{1}{\sigma^2} \left(\frac{1}{\sigma\sqrt{2\pi}} \right) < 0 \quad (\because \sigma > 0).$$

$\therefore x = \mu$ is the mode of the normal distribution.

Median of the normal distribution

Since $f(x)$ is symmetrical about the line $x = \mu$, we have $\int_{-\infty}^{\mu} f(x) dx = \frac{1}{2}$.

Hence μ is the median of X .

Note: If X is a normal variate with mean μ and variance σ^2 then its mean, median and mode coincide.

The points of inflexion of the normal curve are $\mu + \sigma$ and $\mu - \sigma$

Proof: The points of inflexion of the normal curve are given by $f''(x) = 0$ and $f'''(x) \neq 0$.

$$f''(x) = 0 \Rightarrow -\frac{f(x)}{\sigma^2} \left[1 - \frac{(x-\mu)^2}{\sigma^2}\right] = 0.$$

$$\Rightarrow \frac{(x-\mu)^2}{\sigma^2} = 1.$$

$$\Rightarrow (x-\mu)^2 = \sigma^2 \rightarrow x - \mu = \pm \sigma \rightarrow x = \mu \pm \sigma$$

$$\begin{aligned} f'''(x) &= -\frac{f(x)}{\sigma^2} \left[-\frac{2(x-\mu)}{\sigma^2}\right] - \frac{f'(x)}{\sigma^2} \left[1 - \frac{(x-\mu)^2}{\sigma^2}\right] \\ &= \frac{f(x)2(x-\mu)}{\sigma^4} + \frac{1}{\sigma^4} f(x) (x-\mu) \left[1 - \frac{(x-\mu)^2}{\sigma^2}\right] \\ &= (x-\mu) \frac{f(x)}{\sigma^4} \left[3 - \frac{(x-\mu)^2}{\sigma^2}\right] \end{aligned}$$

Hence $f'''(x) \neq 0$ at $x = \mu \pm \sigma$.

Hence the points of inflexion are $\mu + \sigma$ and $\mu - \sigma$.

Mean deviation about mean for Normal distribution is $\sigma\sqrt{2/\pi}$.

$$\text{Mean deviation M.D. (about mean)} = \int_{-\infty}^{\infty} |x - \mu| f(x) dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |x - \mu| e^{-(x-\mu)^2/2\sigma^2} dx$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-z^2/2} dz \quad (\text{Putting } \frac{x-\mu}{\sigma} = z)$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \int_0^{\infty} |z| e^{-z^2/2} dz \quad (\because |z| e^{-z^2/2} \text{ is even})$$

$$= \sigma \sqrt{\frac{2}{\pi}} \int_0^{\infty} z e^{-z^2/2} dz \quad (\because |z| = z \text{ when } z \in [0, \infty])$$

$$= \sigma \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-t} dt \quad (\text{putting } \frac{z^2}{2} = t)$$

$$= \sigma \sqrt{\frac{2}{\pi}} [-e^{-t}]_0^{\infty}$$

$$= \sigma\sqrt{2/\pi}$$

M.g.f. of normal distribution about mean is $e^{\sigma^2 t^2/2}$

$$\begin{aligned} M_{x-\mu}^{(t)} &= E[e^{t(x-\mu)}] = e^{-\mu t} E(e^{tx}) \\ &= e^{-\mu t} M_x^{(t)} = e^{-\mu t} e^{\mu t + (t^2 \sigma^2)/2} \\ &= e^{\mu^2 t^2/2} \end{aligned}$$

M.g.f. of standard normal variate is $e^{t^2/2}$

If X is a normal variate with parameters μ and σ^2 then $M_x^{(t)} = e^{\mu t + (t^2 \sigma^2)/2}$

If $\mu=0$ and $\sigma=1$ for a standard normal variate z , then we have $M_z(t) = e^{t^2/2}$

Moments about mean of normal distribution

The moment generating function about mean is $M_{x=\mu}^{(t)} = e^{\sigma^2 t^2/2}$

$$= 1 + (\sigma^2 t^2/2) + \frac{1}{2!} (\sigma^2 t^2/2)^2 + \dots + \frac{1}{n!} (\sigma^2 t^2/2)^n$$

We know $\mu_r =$ coefficient of $\frac{t^r}{r!}$ in the m.g.f. about mean.

$$\therefore \mu_1 = 0; \mu_2 = \sigma^2; \mu_3 = 0; \mu_4 = 3\sigma^4.$$

Note:1 $\mu_{2n+1} = 0$ for all n .

i.e.; All the odd order moments about mean of the normal distribution is zero.

Note 2: $\mu_{2n} = 1.3.5 \dots (2n-1) \sigma^{2n}$.

$$\mu_{2n} = \text{coefficient of } \frac{t^{2n}}{(2n)!} \text{ in the m.g.f. about mean} = \frac{\sigma^{2n} (2n)!}{2^n n!}$$

$$= \frac{\sigma^{2n}}{2^n n!} [1.2.3.4 \dots (2n-1)2n]$$

$$= \frac{\sigma^{2n}}{2^n n!} [1.3.5 \dots (2n-1)] [2.4 \dots 2n]$$

$$= \frac{\sigma^{2n}}{2^n n!} [1.3.5 \dots (2n-1)] 2^n (1.2.3 \dots n)$$

$$= \sigma^{2n} [1.3.5 \dots (2n-1)]$$

β and r coefficients of normal distribution

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3.$$

$$r_1 = \sqrt{\beta_1} = 0.$$

$$r_2 = \beta_2 - 3 = 0.$$

Theorem: Let X_i with mean μ_i and variance σ_i $i=1,2,\dots,n$ be n independent normal variates.

Let $y = \sum_{i=1}^n a_i x_i$. Then y is a normal variate with mean $\sum_{i=1}^n a_i \mu_i$ and variance $\sum_{i=1}^n a_i^2 \sigma_i^2$.

ie., Linear combination of independent normal variates is also a normal variate.

Proof: $M_y(t) = M_{a_1 x_1} + \dots + a_n x_n^{(t)}$

$$= M_{a_1 x_1}(t) M_{a_2 x_2}(t) \dots M_{a_n x_n}(t) \quad (\because x_i \text{ s are independent})$$

$$= M_{x_1}(a_1 t) M_{x_2}(a_2 t) \dots M_{x_n}(a_n t).$$

$$= \left[e^{\mu_1 a_1 t + \frac{t^2 a_1^2 \sigma_1^2}{2}} \right] \left[e^{\mu_2 a_2 t + \frac{t^2 a_2^2 \sigma_2^2}{2}} \right] \dots \left[e^{\mu_n a_n t + \frac{t^2 a_n^2 \sigma_n^2}{2}} \right]$$

$$= e^{t(a_1 \mu_1 + a_2 \mu_2 + \dots + a_n \mu_n) + t^2/2 [a_1^2 \sigma_1^2 + \dots + a_n^2 \sigma_n^2]}$$

$$= e^{t(\sum a_i \mu_i) + (\sum a_i^2 \sigma_i^2) t^2/2} \text{ which is the m.g.f. of the normal distribution with mean } \sum_{i=1}^n a_i \mu_i$$

and variance $\sum_{i=1}^n a_i^2 \sigma_i^2$.

$\therefore Y$ is a normal variate with mean $\sum_{i=1}^n a_i \mu_i$ and variance $\sum_{i=1}^n a_i^2 \sigma_i^2$.

Area property of Normal distribution

Let X be a normal variate with mean μ and variance σ^2 . The probability that random variable X lies between $X=\mu$ and $X=x_1$ is

$$p(\mu < X < x_1) = \int_{\mu}^{x_1} f(x) dx = \frac{1}{\sigma \sqrt{2\pi}} \int_{\mu}^{x_1} e^{-(x-\mu)^2/2\sigma^2} dx$$

Put $\frac{x-\mu}{\sigma} = z$. When $x = \mu$, $z=0$; when $x=x_1$,

$$\frac{x-\mu}{\sigma} = z_1 \text{ (say)}$$

$$\therefore P(\mu < x < x_1) = P(0 < z < z_1) = \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-z^2/2} dz$$

$$= \int_0^{z_1} \phi(z) dz \text{ where } \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-(z^2/2)} \text{ is the probability function of the standard}$$

normal variate.

$\int_0^{z_1} \phi(z) dz$ is known as the **normal probability integral** and gives the area under standard normal curve between the ordinates $z=0$ and $z=z_1$.

$$\text{In particular } P(\mu - \sigma < x < \mu + \sigma) = \int_{\mu - \sigma}^{\mu + \sigma} f(x) dx$$

$$\text{ie., } P(-1 < z < 1) = \int_{-1}^1 \phi(z) dz = 2 \int_0^1 \phi(z) dz$$

$$= 2 \times 0.3413 = 0.6826 \text{ (From area table)}$$

This implies that 68.26% of area is enclosed within the limits $\mu \pm \sigma$.

$$P(\mu - 2\sigma < x < \mu + 2\sigma) = P(-2 < z < 2) = 2 \int_0^2 \phi(z) dz$$

$$= 0.9544.$$

Hence 95.44% (nearly 95%) of the items lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.

$$P(\mu - 3\sigma < x < \mu + 3\sigma) = P(-3 < z < 3) = 0.9973.$$

Hence 99.73% of the items lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: Of particular importance are the values of $z=1.96$ and $z=2.58$. It can be seen from normal area curve table that $P(-1.96 < z < 1.96) = 0.95$ and $P(-2.58 < z < 2.58) = 0.99$

Thus the probability that X lies between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95 and the probability that X lies within $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$ is 0.99. These observations are extensively used in sampling theory for fixing confidential limits and tests of significance.

Quartile deviation for the normal distribution is $= \frac{2}{3} \sigma$.

Let X be a normal variate with mean μ and variance σ^2 . Let Q_1 and Q_3 be the first and third quartiles respectively.

$$\text{We know that Q.D.} = \frac{Q_3 - Q_1}{2}$$

- By definition $P(X < Q_1) = 0.25$ and $P(X > Q_3) = 0.25$

$$\text{Let } z = \frac{x - \mu}{\sigma}, \quad z_1 = \frac{Q_3 - \mu}{\sigma}. \quad \text{Then } -z_1 = \frac{Q_1 - \mu}{\sigma}$$

$$\therefore \frac{Q_3 - Q_1}{\sigma} = 2z_1$$

$$\therefore \text{Q.D.} = \frac{Q_3 - Q_1}{2} = \sigma z_1$$

$$P(0 < z < z_1) = 0.25 \Rightarrow z_1 = 0.67 \text{ (from table)}$$

$$\therefore \text{Q.D.} = \sigma z_1 = 0.67\sigma = \frac{2}{3}\sigma$$

Fitting of normal distribution

To fit a normal distribution to a given data we first calculate with mean μ and the standard deviation σ . Thus the normal curve fitted to the given data is given by

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

Expected frequencies (by area method)

To calculate the expected normal frequencies we proceed as follows:

- Find the value of the standard normal variates corresponding to the lower limit of each of the class interval. $z_i = \frac{x_i - \mu}{\sigma}$ where x_i is the lower limit of the i^{th} class interval.
- Then $\phi(z_i)$, the areas under the normal curve to the left of the ordinate $z = z_i$ are calculated from the tables.
- The difference in areas = $\Delta\phi(z) = \phi(z_{i+1}) - \phi(z_i)$ are calculated for $i=1, 2, 3, \dots$
- The expected normal frequencies are given by $N\Delta\phi(z)$.

Expected frequencies (by ordinates of the normal curve method)

- Find the standard normal variates $z_i = \frac{x_i - \mu}{\sigma}$ corresponding to the mid point of each of the class interval.
- Find $\phi(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ from the table of ordinates.
- Calculate $y = \frac{N}{\sigma} \phi(z)$.
- The expected frequencies are then obtained by multiplying y by the length of the class intervals.

Example: Fit a normal curve to the following data and calculate the expected frequencies by (i) the area method (ii) by the ordinate method.

Class Intervals	60-62	63-65	66-68	69-71	72-74
f	5	18	42	27	8

Solution: We calculate the mean μ and S.D. σ from the following table.

Mid x_i	Class	frequency f_i	$u_i = \frac{x_i - 67}{3}$	$f_i u_i$	$f_i u_i^2$
61	59.5 - 62.5	5	-2	-10	20
64	62.5 - 65.5	18	-1	-18	18
67	65.5 - 68.5	42	0	0	0
70	68.5 - 71.5	27	1	27	27
73	71.5 - 74.5	8	2	16	32
Total		100		15	97

$$\mu = 67 + \frac{15}{100} \times 3 = 67 + 0.45 = 67.45.$$

$$\sigma^2 = \left[\frac{97}{100} + \left(\frac{15}{100} \right)^2 \right] \times 3^2 = [0.97 + 0.0225] \times 9 = 8.9325$$

$$\therefore \sigma = 2.99.$$

The normal curve to be fitted is

$$f(x) = \frac{N}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$= \frac{100}{2.99 \sqrt{2\pi}} e^{-\frac{(x-67.45)^2}{2 \times 8.9325}}$$

We now find the expected frequencies.

(i) Area method

Class	lower class boundary x^1	$z = \frac{x^1 - \mu}{\sigma}$	$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-z^2/2} dz$	$\Delta \phi(z)$	Expected frequencies. $N \Delta \phi(z)$
59.5-62.5	59.5	- 2.66	0.4961	0.0446	4.46 = 5
62.5-65.5	62.5	- 1.66	0.4515	0.2093	20.93=21
65.5-68.5	65.5	- 0.65	0.2422	0.3790	37.90=38
68.5-71.5	68.5	0.35	0.1368	0.2747	27.47=28
71.5-74.5	71.5	1.35	0.4115	0.0794	7.94=8
74.5 -	74.5	2.36	0.4909		

(ii) Ordinate method

Mid x	$z = \frac{x - \mu}{\sigma}$	$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$	$y = \frac{N}{\sigma} \phi(z)$	Expected frequency = $y \times (\text{length of class intervals})$
61	- 2.16	0.0387	1.29	3.87 = 4
64	- 1.15	0.2059	6.88	20.64=21
67	- 0.15	0.3945	13.19	39.57=40
70	0.85	0.2780	9.29	27.87=28
73	1.86	0.0707	2.36	7.08=7

Properties of the Normal distribution

1. The normal curve is 'bell-shaped' and symmetrical in its appearance.
2. The height of the normal curve is at its maximum at the mean.
3. The curve is asymptotic to the base on either side.
4. Since there is only one maximum point, the normal curve is unimodal, ie., it has only one mode.
5. The points of inflexion ie., the points where the change in curvature occurs are $\mu \pm \sigma$.

6. The normal curve is a continuous one.

7. The first and third quartiles are equidistant from the median.

8. $P(\mu - \sigma < x < \mu + \sigma) = 0.6826$.

$P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.9544$.

$P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.9973$.

Example 1:

If X is normally distributed with zero mean and unit variance find the expectation of X^2 .

Solution:

Since X is normally distributed with zero mean and unit variance the normal curve is

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

$$E(x^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x (x e^{-x^2/2}) dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x d(-e^{-x^2/2})$$

$$= \frac{1}{\sqrt{2\pi}} \left\{ [-x e^{-x^2/2}]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2} dx \right\}$$

$$= \frac{1}{\sqrt{2\pi}} [0 + \sqrt{2\pi}] = 1$$

Example 2:

Find the value of k , mean and variance of the following normal distribution

$$f(x) = k e^{-[(x^2/8) + x + 2]}$$

Solution: $f(x) = k e^{-(1/2)[x^2/4 + 2x + 4]}$

$$= k e^{-1/2[(x^2 + 8x + 16)/4]}$$

$$= k e^{-1/2 \left(\frac{x+4}{2} \right)^2}$$

$f(x)$ is a normal distribution if it is of this form $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 \left(\frac{x-\mu}{\sigma} \right)^2}$ where μ is the mean and σ is the standard deviation.

$$\therefore \sigma = 2 \text{ and } \mu = -4 \text{ and } k = \frac{1}{2\sqrt{2\pi}}.$$

Example 3:

If x is normally distributed with mean 8 and S.D.4, find (i) $P(5 \leq x \leq 10)$ (ii) $P(x \geq 15)$ (iii) $P(X \leq 5)$.

Solution:

$$\text{Standard normal variate } z = \frac{x - \mu}{\sigma}$$

$$\mu = 8$$

$$\sigma = 4$$

$$\therefore z = \frac{x - 8}{4}$$

$$\text{When } x = 5, \quad z = \frac{5 - 8}{4} = -0.75.$$

$$\text{When } x = 10, \quad z = \frac{10 - 8}{4} = 0.50.$$

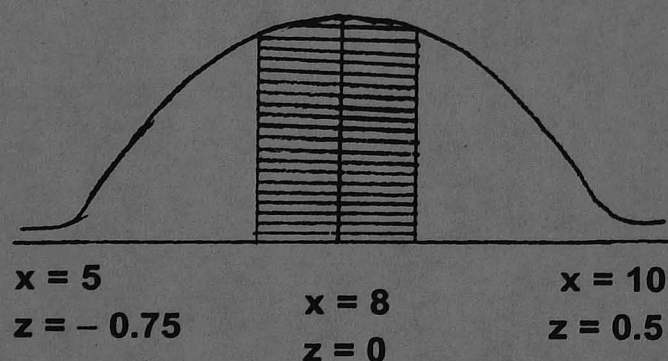
$$\text{When } x = 15, \quad z = \frac{15 - 8}{4} = 1.75.$$

$$\begin{aligned} \text{(i) } P(5 < X < 10) &= P(-0.75 \leq z \leq 0.50) \\ &= P(-0.75 \leq z \leq 0) + P(0 \leq z \leq 0.50) \\ &= P(0 \leq z \leq 0.75) + P(0 \leq z \leq 0.50) \\ &= 0.2734 + 0.1915 = 0.4649. \end{aligned}$$

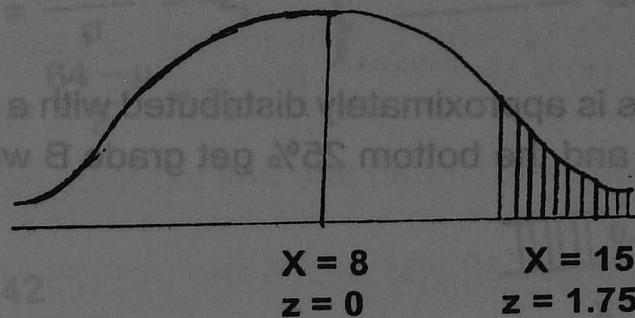
$$\begin{aligned} \text{(ii) } P(x > 15) &= P(z \geq 1.75) = 0.5 - P(0 \leq z \leq 1.75) \\ &= 0.5 - 0.4599 = 0.0401. \end{aligned}$$

$$\begin{aligned} \text{(iii) } P(x < 5) &= P(z \leq -0.75) = 0.5 - P(-0.75 \leq z \leq 0) \\ &= 0.5 - P(0 \leq z \leq 0.75) \\ &= 0.5 - 0.2734 = 0.2266. \end{aligned}$$

Figure for $P(5 \leq x \leq 10)$ is



$$P(x \geq 15)$$



Example 4:

The marks of 1000 students in a university are found to be normally distributed with mean 70 and S.D. 5. Estimate the number of students whose marks will be (i) between 60 and 75. (ii) more than 75 (iii) less than 68.

Solution:

Let X denote the marks of students.

$$\mu = 70$$

$$\sigma = 5.$$

$$\text{The standard normal variate } z = \frac{x - \mu}{\sigma} = \frac{x - 70}{5}$$

(i) To find $P(60 < x < 75)$

$$\text{when } x = 60, \quad z = -2 \text{ and when } x = 75, \quad z = 1.$$

$$\begin{aligned} \therefore P(60 < x < 75) &= P(-2 < z < 1) \\ &= P(-2 < z < 0) + P(0 < z < 1) \\ &= P(0 < z < 2) + P(0 < z < 1) \\ &= 0.4772 + 0.3413 = 0.8185. \end{aligned}$$

$$\therefore \text{The number of students whose marks is between 60 and 75 is } = 1000 \times 0.8185 = 819.$$

(ii) To find $P(x > 75)$,

$$\text{When } x = 75, \quad z = 1.$$

$$\begin{aligned} \therefore P(x > 75) &= P(z > 1) = 0.5 - P(0 < z < 1) \\ &= 0.5 - 0.3413 = 0.1587. \end{aligned}$$

$$\therefore \text{The number of students whose marks is more than 75 is } 159.$$

(iii) To find $P(x < 68)$.

$$\text{When } x = 68; \quad z = \frac{68 - 70}{5} = -0.4$$

$$\begin{aligned} \therefore P(x < 68) &= P(z < -0.4) = P(z > 0.4) \\ &= 0.5 - P(0 < z < 0.4) \\ &= 0.5 - 0.1554 = 0.3546. \end{aligned}$$

$$\therefore \text{The number of students whose marks is less than 68 is } 355.$$

Example 5:

A set of examination marks is approximately distributed with a mean 75 and S.D. of 5. If the top 5% of students get grade A and the bottom 25% get grade B what mark is the lowest A and what mark is the highest B?

Solution:

Let X denote the marks in the examination. Given X is normally distributed with mean $\mu = 75$ and $\sigma = 5$.

Let x_1 be the lowest marks for A and x_2 be the highest marks for B.

Given $P(X > x_1) = 0.05$ and $P(X < x_2) = 0.25$.

The standard normal variate $z = \frac{x_1 - \mu}{\sigma} = \frac{x_1 - 75}{5} = z_1$

$$z = \frac{x_2 - \mu}{\sigma} = \frac{x_2 - 75}{5} = z_2 \quad (1)$$

From the figure we get,

$$P(0 < z < z_1) = 0.45$$

$$\Rightarrow z_1 = 0.45 \text{ (from table)}$$

$$P(-z_2 < z < 0) = 0.25$$

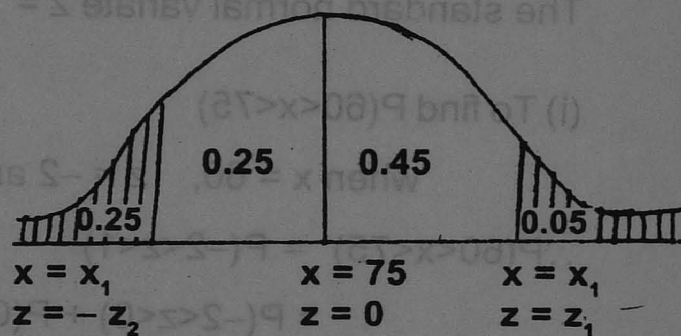
$$\text{ie., } P(0 < z < z_2) = 0.25$$

$$\Rightarrow z_2 = 0.675 \text{ (from the table)}$$

$$\therefore \text{From (1), } x_1 = 75 + 5 z_1 \Rightarrow x_1 = 83.225 = 83.$$

$$x_2 = 75 - 5 z_2 \Rightarrow x_2 = 71.625 = 72.$$

Hence the lowest mark for grade A is 83 and the highest mark for B is 72.



Example 6:

In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation.

Solution:

Let X denote the normal variate with mean μ and standard deviation σ .

$$\text{Given } P(X < 45) = 0.31$$

$$\text{and } P(X > 64) = 0.08$$

$$\left. \begin{aligned} \text{When } X = 45, \quad z &= \frac{45 - \mu}{\sigma} = -z_1 \\ \text{When } X = 64, \quad z &= \frac{64 - \mu}{\sigma} = z_2 \end{aligned} \right\} \quad (1)$$

From the figure

$$P(0 < z < z_2) = 0.42$$

$$\text{and } P(-z_1 < z < 0) = 0.19$$

From the area table we get $z_1 = 0.496$ and $z_2 = 1.405$

$$\text{From (1), } 45 - \mu = 0.496 \sigma$$

$$64 - \mu = 1.405 \sigma$$

Solving the two equations we get $\sigma = 9.99 = 10$; $\mu = 49.96 = 50$.

Example 7:

For a certain normal distribution the first moment about 10 is 40 and the fourth moment about 50 is 48. Find the mean and S.D. of the distribution.

Solution:

$$\text{Given } \mu_1' \text{ (about } x = 10) = 40.$$

$$\mu_4' \text{ (about } x = 50) = 48.$$

$$\text{We know that mean} = A + \mu_1'.$$

$$\mu_1' \text{ is the first raw moment about } X = A.$$

$$\therefore \text{Mean} = 10 + 40 = 50.$$

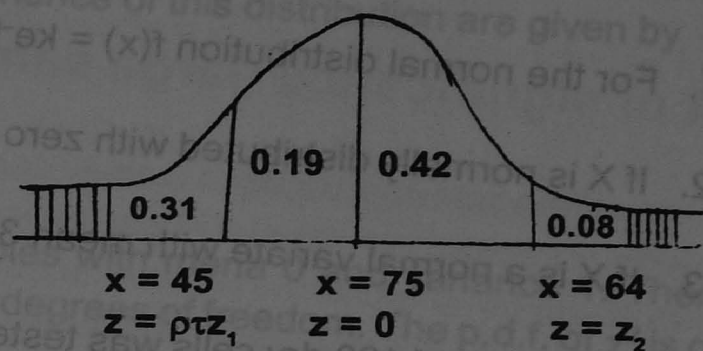
Since mean is 50, we have

$$\mu_4' \text{ (about } x = 50) = \mu_4 = 48.$$

But in a normal distribution, $\mu_4 = 3\sigma^4$.

$$\therefore 3\sigma^4 = 48.$$

$$\therefore \sigma^4 = 16. \quad \therefore \sigma = 2.$$



Exercise

1. For the normal distribution $f(x) = ke^{-(9x^2-12x+13)}$ find k , μ , σ .
2. If X is normally distributed with zero mean and unit variance find the variance of x^2 .
3. If X is a normal variate with mean 30 and S.D. 5 find (i) $P(26 \leq x \leq 40)$ (ii) $P(x > 45)$.
4. A sample of 120 dry cells was tested to find their length of life. The results were $\mu = 11.35$ hrs and $\sigma = 3.03$ hrs. Assuming that the data are normally distributed find the probability that a sample of cells will have a life (i) less than 8 hrs (ii) more than 16 hrs (iii) between 9 and 16 hrs.
5. The average test marks in a particular class is 79. The standard deviation is 5. If the marks are distributed normally, how many students in a class of 200 did not receive marks between 75 and 82.
6. For a normal distribution with mean 50 and standard deviation 15 find Q_1 , and Q_3 .
7. A normal distribution has 77 as mean. Find its S.D. if 20% of the area under the curve lies to the right of 90.
8. At a certain examination 10% of the student who appeared for the paper in Statistics got less than 30 marks and 97% of the students got less than 62 marks. Assuming the distribution to be normal find the mean and S.D.
9. For a normal distribution with mean 2 and S.D. 3. Find a value of the variate such that the probability of the variate from the mean to that value is 0.4115.
10. Fit a normal curve to the following data.

Class	0-20	20-40	40-60	60-80	80-100	Total
Frequency	15	20	20	10	5	70

7.4 Gamma distribution

The Gamma function Γ is defined by $M(p) = \int_0^{\infty} x^{p-1} e^{-x} dx$, $p > 0$.

Hence the function $f(x)$ defined by $f(x) = \begin{cases} \frac{\alpha^\lambda x^{\lambda-1} e^{-\alpha x}}{\Gamma(\lambda)} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$

the probability density function. This distribution is called the Gamma distribution where α and λ are parameters of this distribution. The mean and variance of this distribution are given by $\frac{\lambda}{\alpha}$ and $\frac{\lambda}{\alpha^2}$.

7.5. - Chi- Square distribution

Let $X_1, X_2, \dots, X_n$ be independent normal variables with mean 0 and variance 1. Then the random variable $\chi^2(n) = X_1^2 + \dots + X_n^2$ is called χ^2 with n degrees of freedom. The p.d.f. of χ^2 is given by

$$f(x) = \begin{cases} \frac{1}{2^{n/2} \Gamma(n/2)} x^{n/2-1} e^{-x/2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$f(x)$ can be obtained from the probability density function of the M variable by taking $\lambda = \frac{n}{2}$. The mean and variance of $\chi^2(n)$ are given by n and $2n$.

7.6.- Student's t - distribution

Let $X, X_1, X_2, \dots, X_n$ be $n+1$ independent random variables with parameters 0 and σ .

$$t = \frac{x}{\sqrt{\sum_{i=1}^n X_i^2}}$$

The p.d.f. of the random variable t is given by $f(t) = \frac{1}{\sqrt{n\pi} \Gamma(n/2)} \frac{\Gamma(\frac{n+1}{2})}{(1 + t^2/n)^{\frac{n+1}{2}}}$, $-\infty < t < \infty$

This distribution is called Student's t - distribution with n degrees of freedom.

7.7. - Snedecor's F - distribution

Let u and v be χ^2 variates with m and n degrees of freedom respectively. Then the p.d.f. of the random variable $X = \frac{u}{v}$ is given by

$$f(x) = \frac{(m/n)^{m/2} x^{(m/2)-1} \Gamma[(m+n)/2]}{\Gamma(m/2) \Gamma(n/2) [1 + (mx/n)]^{(m+n)/2}} \text{ if } x > 0$$

and 0 otherwise is called the F - distribution with degrees of freedom (m, n) .

7.8. - Z - distribution

Let X be a random variable having F -distribution with degrees of freedom (m, n) . Then the p.d.f. of the random variable $z = \frac{1}{2} \log_e x$ is given by

$$f(x) = \frac{2(m/n)^{m/2} e^{mz}}{\Gamma(m/2) \Gamma(n/2)} \frac{\Gamma(\frac{m+n}{2})}{(1 + \frac{m}{n} e^{2z})^{m+n/2}}$$

This distribution is called Z - distribution.

Unit - 8

Test of significance (large samples)

Any statistical investigation usually deals with the study of some characteristics of a collection of objects. For example the statistical investigation of estimating the average per capita (monthly) income of the people of India deals with all the earning individuals in the country. Such a group of objects under study is called **population or universe**. It is obvious that for any statistical investigation, complete enumeration of the entire population is practically not possible. This leads to the concepts of sampling.

8.1: Sampling:

Definition: A finite subset of a population is called a **sample** and the number of objects in a sample is called the **sample size**.

Sample error: In order to determine some population characteristics, the objects in the sample are observed and the sample characteristics are used to approximately estimate the same for the entire population. The inherent and unavoidable error in any such approximation is known as **sample error**.

Sampling is quite common in day-to-day life. For example, examining the sample of a particular stuff we decide whether to purchase the same or not. A housewife tests cooked products by examining a small sample by taking a spoonful. Some of the important types of sampling are

(i) Purposive sampling

(ii) Random sampling

(iii) Simple sampling

(iv) Stratified sampling

(i) **Purposive sampling:** If the sample elements are selected with a definite purpose in mind then the sample selected is called purposive sample. Purposive sampling yields favouritism and nepotism in the selection of individuals in the sample. Hence this type of sampling does not fully represent the population. For example, if we want to degrade a particular brand of commodity we may consider the individuals who have complaints of the brand and safely ignoring the non complaints.

(ii) **Random Sampling:** A random sample is one in which each element of the population has an equal chance of inclusion in the sample. Sampling where each member of a population may be chosen more than once is called sampling with replacement while if any member cannot be chosen more than once it is called sampling without replacement.

- (iii) **Simple sampling:** Simple sampling is a special type of random sampling in which each element of the population has an equal and independent chance of being included in the sample.

Simple sampling is a random sampling but a random sampling is not necessarily a simple sampling. For example consider a random sample of K balls chosen from an urn containing 'a' white balls and 'b' black balls. The probability of drawing a white ball in the first draw is $\frac{a}{a+b}$ and if this ball is not replaced, the probability of drawing a white ball in the second draw is $\frac{a-1}{a+b-1}$. Hence the sampling is not simple. But in the case of finite population random sampling with replacement is a simple sampling and in the case of infinite population any random sampling is simple.

A sample of size n from a population may be identified with a series of n independent trials with constant probability p of success for each trial.

- (iv) **Stratified sampling:** The sample which is the aggregate of the sampled individuals of each stratum is called stratified sample and the technique of selecting such sample is called stratified sampling. For example, the population of employees may be divided into two strata-rural employees, urban employees. After dividing the population into strata we select individuals at random from each stratum. Stratified sample is considered to be the best and it is a good representative of the population from which it has been drawn.

8.2 Sampling distribution

Statistical measures computed from a sample of the population are called statistic. For example, if μ is the mean of the population then $\bar{x}$ is the sample mean. If σ^2 is the variance of the population then s^2 is the sample variance. $\bar{x}$ and s^2 are statistics of the population.

Consider all possible samples of size n from a given finite population of size N . For any sample $\{x_1, x_2, \dots, x_n\}$ we can compute some statistic such as $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, the variance $s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$. In general, let $t(x_1, x_2, \dots, x_n)$ be a statistic of the sample. The set of all such statistics, one for each sample, is called the **sampling distribution** of the statistics. For example the mean and variance of the sampling distribution of the statistic t are given by $\bar{t} = (t_1 + t_2 + \dots + t_k)/k$ and $\text{var } t = \sum (t_i - \bar{t})^2/k$.

Example: Consider the collection of 25 samples of size 2 chosen from the population 1,2,3,4,5.

1,1	1,2	1,3	1,4	1,5
2,1	2,2	2,3	2,4	2,5
3,1	3,2	3,3	3,4	3,5
4,1	4,2	4,3	4,4	4,5
5,1	5,2	5,3	5,4	5,5

Population mean $\mu = (1+2+3+4+5)/5 = 3$.

The population variance is $\sigma^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$
 $= 2$.

The sample means of the 25 samples of size 2 are given in the table.

1	1.5	2	2.5	3
1.5	2	2.5	3	3.5
2	2.5	3	3.5	4
2.5	3	3.5	4	4.5
3	3.5	4	4.5	5

Mean of sample means $= \mu_{\bar{x}}$
 $= \frac{\text{Sum of the sample means}}{25} = \frac{75}{25} = 3$.

$\mu = \mu_{\bar{x}} = 3$.

It can be verified that variance of the sampling distribution of means $\sigma_{\bar{x}}^2 = \Sigma(\bar{x}_i - \mu_{\bar{x}})^2/n = 1$.

$\therefore \sigma^2 \neq \sigma_{\bar{x}}^2$.

Standard Error (SE): The standard deviation of the sampling distribution of a statistic is known as **standard error**.

$\mu, \sigma, P, \mu_r, \rho$ respectively denote the population mean, population S.D., proportion of population, r^{th} moment of the population and the correlation coefficient. Correspondingly, $\bar{x}, s, p, m_r, r$ respectively denote the sample mean, sample S.D., proportion of sample, r^{th} moment of the sample and sample correlation coefficient.

Standard Errors for some sampling distributions:

Sampling Statistics	Means $\bar{x}$	Proportions p	S.D. S	Variance S^2	Correlation Coefft. r
Standard	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$\sigma_p = \sqrt{\frac{PQ}{n}}$	$\sigma_s = \frac{\sigma}{\sqrt{2n}}$	$\sigma_s^2 = \sigma^2 \sqrt{2/n}$	$\sigma_r = \frac{1-\rho^2}{\sqrt{n}}$

Note 1: Standard error plays a very important role in the theory of large samples and forms the basis of the testing of hypothesis.

Note 2: If t is any statistic, then for large samples $z = \frac{t - E(t)}{\sqrt{\text{Var } t}}$ with parameters 0 and 1 asymptotically as $n \rightarrow \infty$ and hence $z = \frac{t - E(t)}{\text{S.E.}}$ asymptotically as $n \rightarrow \infty$.

8.3 Testing of hypothesis

Statistical estimation and testing of statistical hypothesis are two important aspects in statistical inference. Statistical estimation is the name given to a technique of estimating the population values of parameter on the basis of information contained in a sample. There are two types of estimates. They are: 1. Point estimates 2. Interval estimates.

- 1. Point estimates:** A point estimate is a single number which is used as an estimate of the unknown population parameter.
- 2. Interval estimates:** An interval estimate of a population parameter is a statement of two values between which it is estimated that the parameter lies.

Estimate: The numerical value of the sample mean is said to be an estimate of the population mean figure.

Estimator: The statistical measure used that is, the method of estimation is referred to as an estimator.

Properties of a good estimator

1. Unbiasedness.
2. Consistency.
3. Efficiency.
4. Sufficiency.

Definition: Hypothesis in statistics is a statement about the values of the population parameter. Testing of hypothesis is a procedure for deciding whether to accept or reject the hypothesis. Procedures which enable us to decide whether to accept or reject hypothesis are called **tests of**

hypothesis also known as **tests of significance**. The test used when we deal with large sample is known as **normal test of significance**.

Null hypothesis: Null hypothesis is a hypothesis tested for possible rejection under the assumption that it is true. It is denoted as H_0 .

For example, to test whether there is a difference between two populations we take the null hypothesis as H_0 ; there is no difference between the two populations. Rejecting this null hypothesis H_0 will mean that the two populations are different. Acceptance of the null hypothesis will mean that there is no difference in the population.

Alternative hypothesis: It is any hypothesis other than the null hypothesis. It is denoted by H_1 .

For example, if we want to test the null hypothesis that the population has specified mean μ_0 then the null hypothesis is $H_0: \mu = \mu_0$. The alternative hypothesis H_1 will be (i) $H_1: \mu \neq \mu_0$. (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 hypothesis in (ii) is known as **right tailed alternative** and the alternative hypothesis in (iii) is known as **left tailed alternative**.

Errors in testing of hypothesis

There are 2 possible types of errors which may arise in testing a statistical hypothesis. I. We may reject a hypothesis H_0 when H_0 is true. II. We may accept a hypothesis H_0 when H_0 is false.

The above two errors are known as Type I error and Type II error respectively.

The test procedure should be so framed as to safeguard against both these types of errors.

Definition: The probability level below which we reject a null hypothesis is called the **level of significance**.

Usually two levels are used namely 5% and 1% level of significance. When we take 5% as the level of significance then the probability of committing Type I error is 0.05. Similarly, if we take 1% as the level of significance then the probability of committing Type I error is 0.01.

Definition: A test of a statistical hypothesis H_0 involves determination of region in the sample space such that the hypothesis will be rejected if the sample point falls within the region and it will be accepted if the sample point falls outside the region. This region is called the **critical region** or **region of rejection**. The rest of the sample space is called the **region of acceptance**.

The value of a test statistic which separates the critical region and the region of acceptance is called the critical value or significant value. It depends on (i) the level of significance (ii) the alternative hypothesis whether it is two tailed or single tailed.

Procedure for testing of hypothesis for large samples

For large samples, we know the standardised variable corresponding to the statistic t is $z = \frac{t - E(t)}{S.E.}$ with mean 0 and variance 1. The value of z under the null hypothesis H_0 is known as the test statistic. Then the critical value of the test statistic at level of significance say α , for a two tailed alternatives is given by z_α where z_α is determined by the equation $P(|z| > z_\alpha) = \alpha$.

Since the standard normal curve is symmetric about the origin we have

$$P(|z| > z_\alpha) = \alpha \rightarrow P(z > z_\alpha) + P(z < -z_\alpha) = \alpha.$$

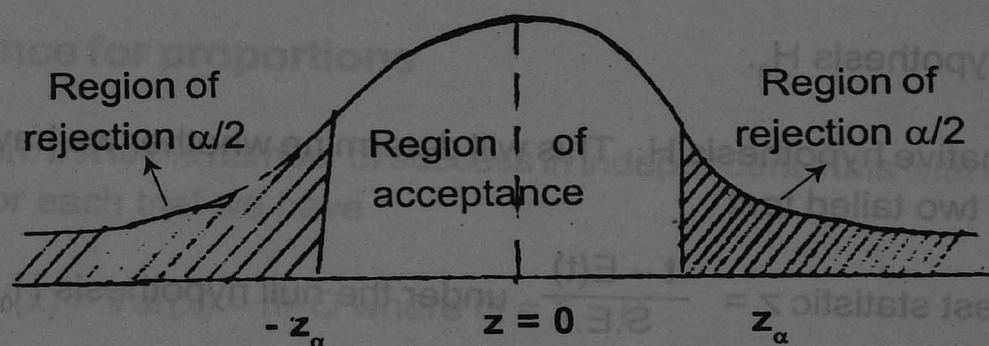
$$\Rightarrow P(z > z_\alpha) + P(z > z_\alpha) = \alpha.$$

$$\Rightarrow 2P(z > z_\alpha) = \alpha$$

$$\Rightarrow P(z > z_\alpha) = \frac{\alpha}{2}$$

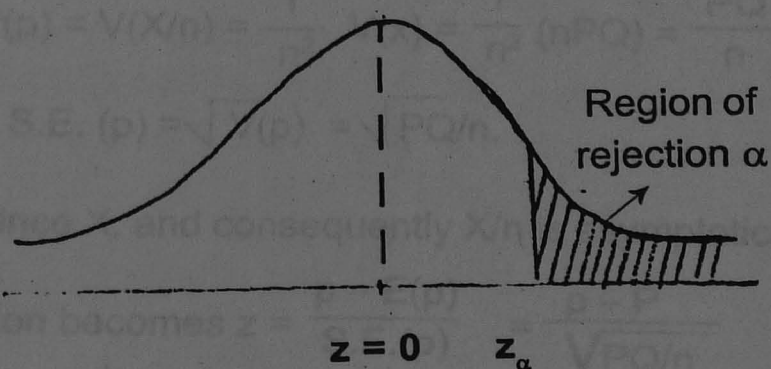
$$\Rightarrow \text{The area of each tail is } \frac{\alpha}{2}$$

Two tailed test (level of significance α)

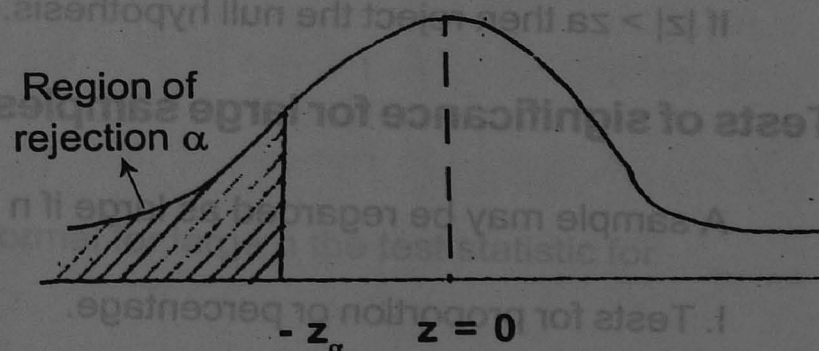


In the case of single tailed alternatives the critical value z_α is determined by the equation $P(z > z_\alpha) = \alpha$ or $P(z < -z_\alpha) = \alpha$ according as the test is right tailed or left tailed.

Right tailed test



Left tailed test



The critical values for different levels of significance are given in the following table.

Critical values z_{α}	Level of significance	
	1%	5%
Two tailed test	$ z_{\alpha} = 2.58$	$ z_{\alpha} = 1.96$
Right tailed test	$z_{\alpha} = 2.33$	$-z_{\alpha} = 1.645$
Left tailed test	$z_{\alpha} = -2.33$	$z_{\alpha} = -1.645$

Note: From the normal probability table we have

$$P(-3 \leq z \leq 3) = 0.9973$$

$$\therefore P(|z| \leq 3) = 0.9973$$

$$\therefore P(|z| > 3) = 1 - P(|z| \leq 3) = 1 - 0.9973 = 0.0027$$

i.e., We can expect surely in all probability a standard normal variate to lie between -3 and 3.

The procedure for testing of a statistical hypothesis

Step I Set up the null hypothesis H_0 .

Step II Set up the alternative hypothesis H_1 . This will determine whether we have to use right tailed test or left tailed test or two tailed test.

Step III Compute the test statistic $z = \frac{t - E(t)}{S.E.}$ under the null hypothesis H_0 .

Step IV Choose the appropriate level of significance α (generally) 5% or 1%.

Step V Compare the computed value of $|z|$ in step IV with the critical value z_{α} at the required level of significance α .

Conclusion: If $|z| < z_{\alpha}$ then accept the null hypothesis.

If $|z| > z_{\alpha}$ then reject the null hypothesis.

Tests of significance for large samples

A sample may be regarded as large if $n > 30$.

I. Tests for proportion or percentage.

(A) Single proportion (B) Difference of proportions.

II. Tests for means.

(A) (i) Test for single mean if standard deviation of the population σ is known i.e. $H_0: \mu = \mu_0$, σ is known.

(ii) Test for single mean if σ is not known $H_0: \mu = \mu_0$, σ is unknown.

(B) (i) Test for equality of means of 2 normal populations with known standard deviations. i.e. $H_0: \mu_1 = \mu_2$; σ_1, σ_2 known.

(ii) Test for equality of means of 2 normal populations with same standard deviations though unknown $H_0: \mu_1 = \mu_2$, given $\sigma_1 = \sigma = \sigma_2$.

III Test for standard deviations

(A) Test for single standard deviation $H_0: \sigma = \sigma_0$.

(B) Test for equality of 2 standard deviations i.e., $H_0: \sigma_1 = \sigma_2$.

IV Test for correlation.

8.4 Test of significance for proportions

(a) Single proportion: If X is the number of success in independent trials with constant probability of success P for each trial we have

$E(x) = np$ and $V(x) = \text{Var}(x) = nPQ$ where $Q = 1 - P$.

$$Z = \frac{X - E(x)}{\text{S.E.}} = \frac{X - nP}{\sqrt{nPQ}} \text{ is a normal variate with mean 0 and variance 1.}$$

$\therefore Z$ is a test statistic and normal test can be applied.

Note: If X is the number of persons, in a sample of size n possessing the given attribute then the observed proportion of success $p = \frac{X}{n}$.

$$\therefore E(P) = E(X/n) = \frac{1}{n} E(X) = n P/n = P.$$

$$V(p) = V(X/n) = \frac{1}{n^2} V(x) = \frac{1}{n^2} (nPQ) = \frac{PQ}{n}.$$

$$\therefore \text{S.E.}(p) = \sqrt{V(p)} = \sqrt{PQ/n}.$$

Since X , and consequently X/n is asymptotically normal for large n the test statistic for

$$\text{proportion becomes } z = \frac{p - E(p)}{\text{S.E.}(p)} = \frac{p - P}{\sqrt{PQ/n}}$$

Under $H_0: p=P$ normal test can be applied.

Since the probable limits for a normal variate X are $E(x) \pm 3\sqrt{\text{Var}(x)}$ the probable limits for the observed proportion of successes are $E(p) \pm 3 \text{ S.E. } (p)$ ie $P \pm 3 \sqrt{PQ/n}$.

If P is not known then the probable limits for the proportion in the population are $p \pm 3\sqrt{pq/n}$.

(B) Difference of proportions: Suppose we want to compare 2 distinct populations, with regard to possession of an attributes. Let a sample of size n_1 be chosen from the first population and another sample of size n_2 be chosen from the second population. Let X_1 be the number of persons possessing the attribute A in the first sample and X_2 be the number of persons possessing the same attribute in the second sample $p_1 = X_1/n_1$

$p_2 = X_2/n_2$. $E(p_1) = P_1$ and $E(p_2) = P_2$ where P_1 and P_2 are the proportions in the populations.

$V(p_1) = P_1Q_1/n_1$ and $V(p_2) = P_2Q_2/n_2$.

Since, for large samples p_1 and p_2 are asymptotically normally distributed $p_1 - p_2$ is also normally distributed and also

$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2$ and

$V(p_1 - p_2) = V(p_1) + V(p_2)$.

$\therefore Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}}$

$= \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$ is a standard normal variate with mean 0 and variance 1.

Under the null hypothesis $H_0: P_1 = P_2 = P$ (hence $Q_1 = Q_2 = Q$) test statistic becomes

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

An unbiased estimate of population proportion P based on both the samples is given by $P = (n_1P_1 + n_2P_2)/(n_1 + n_2)$. Suppose the population proportions P_1 and P_2 are given to be different. ie., $P_1 \neq P_2$. If we want to test whether the difference $P_1 - P_2$ is significant then the test statistic becomes

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\text{S.E. } (p_1 - p_2)}$$

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

Here sample proportions are not given. If we set up a null hypothesis $H_0 : p_1 = p_2$ then test statistic becomes

$$z = \frac{|p_1 - p_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

Example 1: A coin is tossed 144 times and a person gets 80 heads. Can we say that the coin is unbiased one?

Solution: Set the null hypothesis H_0 : the coin is unbiased.

Given $n = 144$.

Probability of getting a head in a toss $P = \frac{1}{2}$

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

Let X = number of successes

= number of getting heads = 80.

$$\therefore Z = \frac{80 - 144\left(\frac{1}{2}\right)}{\sqrt{144\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{80 - 72}{\sqrt{36}} = \frac{8}{6} = 1.33 < 1.96$$

Since $|z| < 1.96$ we accept the hypothesis at 5% level of significance. Hence the coin is unbiased.

Example 2: A die is thrown 10000 times and a throw of 1 or 2 was obtained 4200 times. On the assumption of random throwing do the data indicate an unbiased die?

Solution: P = Probability of getting 1 or 2 = $\frac{2}{6} = \frac{1}{3}$

$$\therefore Q = \frac{2}{3}$$

Given $n = 10000$.

x = the number of success = 4200.

The null hypothesis H_0 : The die is unbiased.

$$\begin{aligned}\therefore z &= \frac{X - np}{\sqrt{npq}} = \frac{4200 - 10000^{(1/3)}}{\sqrt{10000^{(2/9)}}} \\ &= \frac{4200 - 3333.3}{47.14} = \frac{866.7}{47.14} = 18.4.\end{aligned}$$

Since $|z| > 3$, H_0 is rejected and hence the die is biased one.

Example 3: In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?

Solution: Here $n = 600$. $p = \frac{325}{600} = 0.5417$.

Set the null hypothesis. H_0 : The number of smokers and non smokers are equal in the city.

$\therefore P =$ population proportion of smokers in the city $= 0.5$.

$\therefore Q = 0.5$

Alternative hypothesis H_1 : $P > 0.5$ (right tailed test).

$$\begin{aligned}\text{Under } H_0 \text{ we have } z &= \frac{p - P}{\sqrt{PQ/n}} = \frac{0.5417 - 0.5}{\sqrt{0.5 \times 0.5/600}} \\ &= \frac{0.0417}{0.0204} = 2.04.\end{aligned}$$

$z = 2.04 > 1.645$ (under 5% level of significance for right tailed test). Hence the difference is significant and H_0 is rejected at 5% level and H_1 is accepted. Hence we conclude that the majority of men in the city are smokers.

$z = 2.04 < 2.33$ (under 1% level of significance for right tailed test). Hence the difference is not significant at 1% level of significance. Here H_0 may be accepted at 1% level of significance.

Example 4: A random sample of 500 apples was taken from a large consignment of apples and 65 were found to be bad. Find the S.E. of the population of bad ones in a sample of this size. Hence deduce that the percentage of bad apples in the consignment almost certainly lies between 8.5 and 17.5.

Solution: Given $n = 500$

Let $X =$ Number of bad apples in the sample $= 65$.

$$p = \text{proportion of bad apples in the sample} = \frac{65}{500} = 0.13$$

$\therefore q = 0.87$

Here the proportion of bad apples P in the population is not known.

$\therefore$ We can take $P = p = 0.13$; $Q = q = 0.87$ and $n = N = 500$.

$$\text{S.E. of proportion} = \sqrt{PQ/N} = \sqrt{(0.13)(0.87)/500} = 0.015$$

Limits for the proportions of bad apples in the population is $P \pm 3\sqrt{PQ/N} = 0.13 \pm 3(0.015)$

$$= 0.13 + 0.045 = 0.175 \text{ and } 0.085$$

$$= 17.5\% \text{ and } 8.5\%.$$

Example 5: During a country wide investigation the incidence of TB was found to be 1% In a college of 400 strength 5 were reported to be affected whereas in another college of 1200 strength 10 were reported to be affected. Does this indicate any significant difference?

Solution: Given $P = 0.01$

$$\therefore Q = 0.99.$$

$$n_1 = 400, p_1 = \frac{5}{400} = \frac{1}{80} = 0.0125$$

$$n_2 = 1200, p_2 = \frac{10}{1200} = \frac{1}{120} = 0.0083.$$

Set the null hypothesis $H_0: P_1 = P_2$.

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.0125 - 0.0083}{\sqrt{0.01 \times 0.99 \left(\frac{1}{400} + \frac{1}{1200}\right)}}$$

$$= \frac{0.0042}{\sqrt{0.01 \times 0.99 \left(\frac{1}{300}\right)}} = 0.7368 < 1.96.$$

Hence the difference is not significant on 5% level of significance.

Example 6: In a year there are 956 births in a town A of which 52.5% were males while in town A and B combined this proportion in a total of 1406 births was 0.496. Is there any significant difference in the proportion of male births in the two towns?

Solution: Given $n_1 = 956$.

p_1 = proportion of male birth = 0.525.

p_2 and n_2 are not given.

But $n_1 + n_2 = 1406$ so that $n_2 = 450$.

$P = 0.496$, So $Q = 1 - P = 1 - 0.496 = 0.504$.

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \Rightarrow 0.496 = \frac{956 \times 0.525 + 450 \times p_2}{1406}$$

$$\Rightarrow p_2 = 0.434$$

Set the null hypothesis $H_0 : P_1 = P_2$ (no difference of births of males in the towns).

$$\therefore z = \frac{p_1 - p_2}{\sqrt{PQ \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.525 - 0.434}{\sqrt{0.496 \times 0.504 \left(\frac{1}{956} + \frac{1}{450} \right)}}$$

$$= 3.14$$

$$z > 3.$$

Hence there is significant difference in the proportion of male births in the towns A and B.

Example 7: A cigarette manufacturing firm claims that its brand A cigarette outsells its brand B cigarette by 8%. It is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another sample of 100 smokers prefer brand B. Test whether the 8% difference is a valid claim.

Solution: Given $n_1 = 200$, $n_2 = 100$.

Proportion of preference of A of the first sample = $p_1 = \frac{42}{200} = 0.21$.

Proportion of preference of B of the second sample = $p_2 = \frac{18}{100} = 0.18$.

Set up the null hypothesis $H_0 : 8\%$ difference in the sale of 2 brands of cigarette is always there. ie., $H_0 : P_1 - P_2 = 0.08$.

Alternative hypothesis $H_1 : P_1 - P_2 \neq 0.08$ (Two tailed alternative)

Since the population proportion is not given we can find it as $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$P = \frac{42 + 18}{300} = 0.20.$$

$$\therefore Q = 0.8.$$

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{PQ \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.03 - 0.08}{\sqrt{0.2 \times 0.8 \left(\frac{1}{200} + \frac{1}{100} \right)}} = -1.02$$

$$\therefore |z| = 1.02 < 1.96$$

The difference is not significant at 5% level of significance. Hence the null hypothesis may be accepted at 5% level. We can conclude that a difference of 8% in the sale of 2 brands is valid claim.

Example 8: Two samples of sizes 1200 and 900 respectively are drawn from 2 large populations. In the 2 large populations there are 30% and 25% respectively of fair haired people. Test whether these 2 samples will reveal the difference in the population proportions.

Solution: Given $n_1 = 1200$

$$n_2 = 900.$$

P_1 = proportion of fair haired people in the first population (30%) = 0.30.

P_2 = proportion of fair haired people in the 2nd population (25%) = 0.25.

Set up the null hypothesis $H_0 : p_1 = p_2$ (the sample proportions are equal).

$$|z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

$$\therefore |z| = \frac{0.30 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}}} = \frac{0.05}{0.0195} = 2.56.$$

$|z| > 1.96$. Hence the hypothesis is rejected at 5% level. Hence these samples will reveal the difference in the population proportion.

Exercise

1. A coin is tossed 400 times and head turns up 216 times. Discuss whether the coin may be unbiased one.
2. A die is thrown 9000 times and a throw of 3 or 4 observed 3240 times. Show that the die cannot be regarded as an unbiased one.
3. In a sample of 500 people in Tamil Nadu 280 are tea drinkers, the rest are coffee drinkers. Can we assume that both coffee and tea are equally popular in Tamil Nadu at 1% level of significance?
4. In a sample of 1000 people in Tamil Nadu 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance.
5. In a sample of 600 students of a certain college 400 are found to use ball pens. In another college from a sample of 900 students 450 were found to use ball pens. Test whether the 2 colleges are significantly different with respect to the habit of using ball pens.

6. An advertising company claims that 40% of the people who saw an advertisement screened on their television set remembered the name of the product advertised 24hrs after they have seen the show. In a sample survey conducted 24 hrs after the show 152 out of 400 persons remembered the name of the product advertised. Test if the claim of the company can be accepted at a level of 1%.
7. Random samples of 400 men and 600 women in a locality were asked whether they would like to have a bus stand near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposals are same in the male and female. Discuss at 5% level of significance.
8. 20 people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate if attacked by this disease is 85% in favour of the hypothesis that it is more at 5% level?
9. 10 dice are thrown 300 times each and a throw of an odd number is reckoned as success. Are the dice unbiased if 1650 throws of odd numbers have been obtained?
10. One make of a motor car developed engine trouble in 5 races out of total of 100 races and another make of motor car of the same year developed engine trouble in 7 races out of a total of 200 races. Find out if there is a significant difference in the two types of motor cars as far as engine defects are concerned.

8.5 Test of significance for means

(A) Test of significance for single mean

(i) Test of significance for single mean if σ is known

If $x_1, x_2, \dots, x_n$ is a random sample of size n from a normal population with mean μ and s.d. σ , then the sample mean $\bar{x}$ is distributed normally with mean μ and s.d. $\sigma/\sqrt{n}$. Thus the standard normal variate corresponding to $\bar{x}$ is $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Set up a null hypothesis $H_0 : \bar{x} = \mu$

(ii) If the s.d. of the population is not known

We can use its estimate of the sample S.D. as $\sigma = s$ and the test statistic is $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

Note: Confidence limits for μ .

The 95% confidence limits for μ is got from $|z| \leq 1.96$.

$$\text{i.e. } \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \Rightarrow \bar{x} - 1.96 (\sigma / \sqrt{n}) \leq \mu \leq \bar{x} + (1.96) \sigma / \sqrt{n}$$

$\therefore \bar{x} \pm 1.96 (\sigma/\sqrt{n})$ are the 95% confidence limits for μ .

Similarly $\bar{x} \pm 2.33 (\sigma/\sqrt{n})$ are the 98% confidence limits for μ and $\bar{x} \pm 2.58 (\sigma/\sqrt{n})$ are the 99% confidence limits for μ .

Example 1: A normal population has a mean of 6.48 and S.D. of 1.5. In a sample of 400 members mean is 6.75. Is the difference significant?

Solution: Given $\mu = 6.48$, $\sigma = 1.5$; $n=400$; $\bar{x} = 6.75$. Let $H_0 : \mu = \bar{x}$. Under this hypothesis the test

$$\text{statistic is } z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{6.75 - 6.48}{1.5 / \sqrt{400}} \\ = \frac{0.27 \times 20}{1.5} = \frac{5.4}{1.5} = 3.6 > 3.$$

$\therefore$ The difference is significant.

Example 2: An insurance agent has claimed that the average age of policy holders who insure through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy holders who had insured through him reveal that the mean and S.D. are 28.8 years and 6.35 years respectively. Test his claim at 5% level of significance.

Solution: Given $\mu = 30.5$; $n=100$; $\bar{x}=28.8$ $s = 6.35$

Set the null hypothesis $H_0 : \mu = 30.5$.

The alternative hypothesis $H_1 : \mu < 30.5$ (left tailed test)

$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{28.8 - 30.5}{6.35 / \sqrt{100}} = \frac{-1.7 \times 10}{6.35} = -2.68$$

$$\therefore |z| = 2.68 > 1.645.$$

It is significant at 5% level of significance. Hence H_0 is rejected and H_1 is accepted at 5% level. Hence the claim of the agent is valid.

Example 3: An entrance examination test was given to a large group of boys and girls who appeared for the selection of professional course in Tamil Nadu who scored on the average 84.5 marks. The same test was given to a small group of 400 boys and girls. They scored an average of 82.5 marks with a S.D. 22.5 marks. Examine whether the difference is significant. Can we conclude that the sample is from the population?

Solution: Given $\mu = 84.5$; $n = 400$; $\bar{x} = 82.5$ and $s = 22.5$.

Here the S.D. of the population is not given.

Let $H_0 : \mu = x$. Under this H_0 the test statistic is

$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{82.5 - 84.5}{22.5 / \sqrt{400}} = \frac{-2 \times 20}{22.5} = -1.778.$$

$\therefore$ The difference is not significant on 5% level of significance. Hence we can conclude that the sample is from the population.

Example 4: A sample of 900 men is found to have a mean height of 64 cm. If this sample has been drawn from a normal population with S.D. 20 cm, find the 99% confidence limits for the mean height of the men in the population.

Solution: Given $n = 900$; $\bar{x} = 64$; $\sigma = 20$.

The 99% confidence limits for the population mean μ is $\bar{x} + 2.58 (\sigma/\sqrt{n})$ and $\bar{x} - 2.58(\sigma/\sqrt{n})$

ie., $64 + 2.58(20/\sqrt{900})$ and $64 - 2.58(20/\sqrt{900})$

ie., $64 + (51.6/30)$ and $64 - (51.6/30)$ ie., $64 + 1.72$ and $64 - 1.72$.

$\therefore$ The 99% confidence limits for the mean height is (62.28, 65.72).

Exercise

1. A sample of size 400 was drawn and the sample mean 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.
2. The medical inspection of a large number of school children revealed an average chest girth of 84.5 cm with a S.D. of 6.4 cm. In a school 400 strong the average chest girth of the children was found to be 82 cm. Is the difference significant?
3. A random sample of 400 items is found to have mean of 82 and S.D. of 18. Find 95% confidence limits for the mean of the population from which the sample is drawn.
4. In a sample survey of 50 individuals the average income was found to be Rs.2,172 per annum. It is known that the S.D. of the distribution of income is 90.11. Can we conclude that average annual income is Rs.2,200.
5. A lady stenographer claims that she can take dictation at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trials in which she demonstrates a mean of 116 words with a S.D. of 15 words?
6. A sample of 5000 students from MKU was taken and their average weight was found to be 62.5 kgs with a S.D. of 22 kgs. Find the 95% confidential limits of the average weight of the students in the entire University.
7. A sample of 900 members has a mean 3.4 cm and S.D. 2.61 cms. Is the sample from a large population of mean 3.25 cms. and S.D. 2.61 cms? If the population is normal and its mean is unknown find the 95% and 98% fiducial limits of true mean?

(B) Test of significance for difference of sample means

Consider two different normal populations with mean μ_1 and μ_2 and s.d. σ_1 and σ_2 respectively. Let a sample of size n_1 be drawn from the first population and an independent sample of size n_2 be drawn from the second population. Let $\bar{x}_1$ be the mean of the first sample from the first population and $\bar{x}_2$ be the mean of the second sample from the second population. If the sample sizes are large we know $\bar{x}_1$ is a normal variate with mean μ_1 and variance σ_1^2/n_1 and $\bar{x}_2$ is an independent normal variate with mean μ_2 and variance σ_2^2/n_2 .

$\therefore$ The difference $\bar{x}_1 - \bar{x}_2$ of the two independent normal variate $\bar{x}_1$ and $\bar{x}_2$ is also an independent normal variate. Hence the standard normal variate corresponding to $\bar{x}_1 - \bar{x}_2$ is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{\text{S.E.}(\bar{x}_1 - \bar{x}_2)}$$

Set the hypothesis $H_0 : \mu_1 = \mu_2$.

Then we get $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2 = 0$ and

$$V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2) = (\sigma_1^2/n_1) + (\sigma_2^2/n_2)$$

$$\begin{aligned}\therefore \text{S.E.}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\text{Var}(\bar{x}_1 - \bar{x}_2)} \\ &= \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}\end{aligned}$$

Thus under H_0 , the test statistic becomes $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$ which can be tested at any level of significance.

Note 1: If the samples have been drawn from the distinct populations with common s.d. σ then under the null hypothesis the above test statistic becomes $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(1/n_1) + (1/n_2)}}$ is a standard normal variate with mean 0 and variance 1 which can be tested at any level of significance.

Note 2: If the common s.d. σ is not known then its estimate based on the sample variances s_1^2 and s_2^2 can be obtained as $\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}$. In this case the test statistic becomes

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Note 3: If $\sigma_1^2 \neq \sigma_2^2$ and σ_1 and σ_2 are not known then they can be estimated from the sample variances as $\sigma_1^2 \sim s_1^2$ and $\sigma_2^2 \sim s_2^2$. In this case the test statistic becomes

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \text{ which can be tested at any required level of significance.}$$

Example 1: The number of accidents per day were studied for 144 days in Madras city and for 100 days in Delhi city. The mean numbers of accidents and the s.d.s were respectively 4.5 and 1.2 for Madras city and 5.4 and 1.5 for Delhi city. Is Madras city more prone to accidents than Delhi city.

Solution: Given $n_1 = 144$; $\bar{x}_1 = 4.5$ $\bar{x}_2 = 5.4$.
 $n_2 = 100$ $\sigma_1 = 1.2$ $\sigma_2 = 1.5$

Set the null hypothesis $H_0 : \mu_1 = \mu_2$.

Since the population standard deviations are not given we take the sample s.d.s as the population s.d and apply the test statistics.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} = \frac{4.5 - 5.4}{\sqrt{(1.2)^2/144 + (1.5)^2/100}} = -4.99.$$

$\therefore |Z| = 4.99 > 3$. we reject the hypothesis that the two cities have the same accident rates. Since Delhi city has higher rate of accidents than Madras city we can say that Delhi is more prone to accidents.

Example 2: The mean yields of rice from two places in a district were 210 kgs and 220 kgs. per acre from 100 acres and 150 acres respectively. Can it be regarded that the samples were drawn from the same district which has the s.d. of 11 kgs per acre.

Solution: $n_1 = 100$; $\bar{x}_1 = 210$; $\sigma = 11$, $n_2 = 150$; $\bar{x}_2 = 220$. Set the null hypothesis $H_0 : \mu_1 = \mu_2$.

$\therefore H_1 : \mu_1 \neq \mu_2$.

$\therefore$ The test statistic $z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{(1/n_1) + (1/n_2)}}$

$$z = \frac{210 - 220}{11 \sqrt{(1/100) + (1/150)}} = \frac{-10}{11 \sqrt{250/15000}} = -7.04.$$

$|z| = 7.04 > 3$. The value is highly significant and hence we reject the null hypothesis. Hence we conclude that the samples are certainly not from the same district with the s.d. 11.

Example 3: A man buys 50 electric bulbs of 'Philips' and 50 electric bulbs of 'HMT'. He finds that 'Philips' bulbs give an average life of 1500 hours with a standard deviation of 60 hours and 'HMT' bulbs gave an average life of 1512 hours with a standard deviation of 80 hours. Is there a significant difference in the mean life of the two makes of bulbs?

Solution: Let us set up the hypothesis that there is no significant difference in the mean life of the two makes of bulbs.

$$\text{Given } n_1 = 50 ; \bar{x}_1 = 1500 ; s_1 = 60$$

$$n_2 = 50 ; \bar{x}_2 = 1512 ; s_2 = 80.$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} = \frac{-12}{\sqrt{(60)^2/50 + (80)^2/50}} = \frac{-12}{14.14} = -0.849.$$

$$|z| = 0.849 < 3.$$

Hence the difference is not significant.

Exercise

1. Intelligence test on two groups of boys and girls gave the following results:

	Mean	S.D.	N
Girls	75	15	150
Boys	70	20	250

Is there a significant difference in the mean scores obtained by boys and girls.

2. Two samples of 100 electric bulbs each has a mean 1500 and 1550. S.D. 50 and 60 respectively. Can it be concluded that two brands differ significantly at 1% level of significance in equality.
3. A college conducts both day and night classes intended to be identical. A sample of 100 day students yields examination result, mean 72.4 and S.D.=14.8. A sample of 200 night students yields examination result mean 73.9 and S.D. 17.9. Are the two means statistically equal at 5% level.?
4. The means of simple samples of 1000 and 2000 are 67.5 and 68.8 inches respectively. Can the samples be regarded as drawn from the same population of s.d. 2.5 inches.
5. A simple sample of heights of 6400 English men has a mean of 67.85 inches and S.D. 2.56 inches while a simple sample of heights of 1600 Australians has a mean of 68.55 and a S.D. of 2.52 inches. Do the data indicate that Australians are on the average taller than English men?

8.6 Test of significance for Standard deviation

(A) Test of significance for single s.d.

If we want to test whether a sample with known standard deviation s could have come from a population with S.D. σ we can use the test statistic as $z = \frac{s - \sigma}{\sigma/\sqrt{2n}}$ under the null hypothesis

$H_0 : \sigma = s$. For it has been found that the standard error of s is approximately $\sigma/\sqrt{2n}$ for large n .

(B) (i) Test of significance for equality of Standard deviations of 2 normal populations

Consider two normal populations with S.D. σ_1 and σ_2 respectively. Let two independent random samples of large sizes n_1 and n_2 having standard deviations s_1 and s_2 be drawn from the first and second normal populations respectively.

Under the null hypothesis $H_0 : \sigma_1 = \sigma_2$ the test statistic is $z = \frac{s_1 - s_2}{\sqrt{(\sigma_1^2/2n_1) + (\sigma_2^2/2n_2)}}$ which

is a standard normal variate with mean 0 and variance 1 which can be tested at any level of significance.

(ii) Test of significance for equality of Standard deviations of a normal population

If we want to test whether the two independent samples with known standard deviations s_1 and s_2 have come from the same population with standard deviation σ . Under the hypothesis

$H_0 : \sigma_1 = \sigma_2$ the test statistic is $z = \frac{s_1 - s_2}{\sigma \sqrt{(1/2n_1) + (1/2n_2)}}$

Example 1: A manufacturer of electric bulbs, by some process, finds the s.d. of the life of the lamps to be 100 hrs. He wants to change the process if the new process results in even a smaller variation in the life of lamps. In adopting the new process a sample of 150 bulbs gave the S.D. of 95 hrs. Is the manufacturer justified in changing the process?

Solution: Given $\sigma = 100$; $s = 95$; $n = 150$.

Set the null hypothesis $H_0 : \sigma = s$.

$$z = \frac{s - \sigma}{\sigma/\sqrt{2n}} = \frac{95 - 100}{100/\sqrt{300}} = \frac{-5\sqrt{300}}{100}$$

$$= \frac{-5 \times 17.32}{100} = \frac{-86.6}{100} = -0.866$$

$$|z| = 0.866 < 1.96.$$

Hence the deviation is not significant at 5% level and hence the hypothesis is accepted. So the manufacturer finds no justification in changing the process on this evidence alone.

Example 2: In a sample of 1000 the mean is 17.5 and the s.d. 2.5. In another sample of 800 the mean is 18 and s.d. 2.7. Assuming that the samples are independent discuss whether the two samples can have come from a population which have the same s.d.

Solution: Let us take the hypothesis that there is no significant difference in the standard deviations of the two samples.

$$\text{Given } n_1 = 1000 ; s_1 = 2.5$$

$$n_2 = 800 ; s_2 = 2.7$$

$$z = \frac{s_1 - s_2}{\sqrt{(\sigma_1^2/2n_1) + (\sigma_2^2/2n_2)}}$$

$$= \frac{2.5 - 2.7}{\sqrt{\frac{(2.5)^2}{2000} + \frac{(2.7)^2}{1600}}} = \frac{-0.2}{0.0876} = -2.283.$$

$$|z| = 2.283 > 1.96.$$

∴ The hypothesis is rejected. Hence the two samples have not come from a population which has the same standard deviation.

Example 3: The s.d. of weight of all students in a first grade college was found to be 4 kgs. Two samples are drawn. The s.d.s. of the weight of 100 undergraduate students is 3.5 kgs and 50 post graduate students is 3 kgs. Test the significance of the difference of standard deviations of the samples at 5% level.

Solution: Given $n_1 = 100 ; s_1 = 3.5 ; \sigma = 4$.

$$n_2 = 50 ; s_2 = 3.$$

Set the null hypothesis $H_0 : \sigma_1 = \sigma_2$.

Then $H_1 : \sigma_1 \neq \sigma_2$.

$$\therefore z = \frac{s_1 - s_2}{\sigma \sqrt{(1/2n_1) + (1/2n_2)}} = \frac{3.5 - 3}{4 \sqrt{1/200 + 1/100}} = 1.02.$$

$$z = 1.02 < 1.96.$$

It is not significant at 5% level of significance.

Exercise

1. The S.D. of a random sample of 900 members is 4.6 and that of another independent sample of 1600 is 4.8 Examine if the S.D. are significantly different.

2. The S.D. of a random sample of 1000 is found to be 2.6 and the S.D. of another random sample of 500 is 2.7. Assuming the samples to be independent discuss whether the two samples could have come from the universe with the same S.D.
3. From the experience of a manufacturer of battery cells according to a technique the S.D. of the life of battery cells was 150 days. He is interested to introduce a new technique in manufacturing batteries with less variation in life of batteries. In his new technique with a sample of 200 batteries he got the S.D. of the life of battery cells as 140 days. Is the manufacturer justified in changing the technique?
4. The mean produce of wheat of a sample of 100 fields is 200 lb. per acre with a S.D. of 10 lb. Another sample of 150 fields gives the mean at 220 lb. with a S.D. of 12 lb. Assuming the S.D. of the mean field at 11 lb for the universe, find at 1% level if the two results are consistent.

8.7 Test of significance for correlation coefficient

The sampling distribution of the correlation coefficient r for random sample of size n from a correlated bivariate normal population with correlation coefficient ρ is far from normal as this probability curve is very skew in the neighbourhood of $\rho = \pm 1$ even for large n . R.A. Fisher showed

that the transformed variable $z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$ is distributed normally with mean

$$z_0 = \frac{1}{2} \log_e \left(\frac{1+\rho}{1-\rho} \right) \text{ and variance } \frac{1}{n-3}.$$

Given a random sample of size n from the bivariate normal population to test the hypothesis

$H_0: r = \rho$ we take the test statistic as $z = \frac{z - z_0}{1/\sqrt{n-3}}$ is a standard normal variate with mean 0 and var 1. and proceed in the usual way.

The same transformation can be used for testing the equality of two population correlations. Let n_1, r_1, ρ_1 be the sample size, sample correlation coefficient and population correlation coefficient for the first population and let n_2, r_2, ρ_2 be for the second population.

To test the null hypothesis $H_0: \rho_1 = \rho_2$ we calculate $z_1 = \frac{1}{2} \log_e \left(\frac{1+r_1}{1-r_1} \right)$ and

$$z_2 = \frac{1}{2} \log_e \left(\frac{1+r_2}{1-r_2} \right)$$

Since z_1 and z_2 are normally distributed their difference $z_1 - z_2$ is also normally distributed with mean zero. Also variance of $z_1 - z_2 = \frac{1}{n_1-3} + \frac{1}{n_2-3}$.

$\therefore$ To test the hypothesis $H_0: \rho_1 = \rho_2$ we take the test statistic as

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \text{ which is a std normal variate with mean 0 and var 1.}$$

Example 1: In a random sample of 50 pairs of values the correlation was found to be 0.89. Is this consistent with the assumption that the correlation in the population is 0.84?

Solution: Given $n = 50$; $r = 0.89$; $\rho = 0.84$.

Let the null hypothesis be $H_0 : r = \rho$.

$$z = \frac{z_1 - z_0}{1/\sqrt{n-3}}$$

$$z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right) = \frac{1}{2} \log_e \frac{1+0.89}{1-0.89}$$

$$= \frac{1}{2} \log_e \left(\frac{1.89}{0.11} \right)$$

$$= \log_e 4.1450 = \log_{10} 4.1450 \times \log_e 10$$

$$= 0.618 \times 2.3026 = 1.42.$$

$$z_0 = \frac{1}{2} \log_e \left(\frac{1+\rho}{1-\rho} \right) = \frac{1}{2} \log_e \frac{1+0.84}{1-0.84}$$

$$= \frac{1}{2} \log_e \left(\frac{1.84}{0.16} \right)$$

$$= \log_e 3.3912 = 0.5304 \times 2.3026 = 1.22.$$

$$\therefore z = \frac{1.42 - 1.22}{\sqrt{1/47}} = 0.20\sqrt{47} = 1.37.$$

$\therefore |z| = 1.37 < 1.96$. Hence the difference is not significant at 5% level of significance. Hence the hypothesis is accepted.

Example 2: A sample of 67 boys was taken and Karl Pearson's coefficient of correlation between two attributes x and y was found to be 0.72. Another sample of 39 girls was taken and coefficient of correlation between the same attributes was found to be 0.84. Can these two samples be considered as coming from populations having equal correlation coefficients?

Solution: Given $n_1 = 67$; $n_2 = 39$; $r_1 = 0.72$; $r_2 = 0.84$.

$$z_1 = \frac{1}{2} \log_e \left(\frac{1+0.72}{1-0.72} \right) = \frac{1}{2} \log_e \left(\frac{1.72}{0.28} \right)$$

$$= 0.9079.$$

$$z_2 = \frac{1}{2} \log_e \left(\frac{1+0.84}{1-0.84} \right) = \frac{1}{2} \log_e \left(\frac{1.84}{0.16} \right) = 1.221.$$

Under the null hypothesis $H_0 : \rho_1 = \rho_2$ the test statistic is

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0.9079 - 1.221}{\sqrt{\frac{1}{64} + \frac{1}{64}}} = \frac{-0.3131 \times 8 \times 6}{10} = -1.503.$$

$$\therefore |z| = 1.503 < 1.96.$$

Hence the difference is not significant on 5% level of significance. Hence H_0 is accepted.

Exercise

1. The correlation coefficient between temperature of rice and breakage percentage calculated from 2 samples 120 and 160 are 0.69 and 0.74 respectively. Do the two estimates differ significantly.
2. The Pearson coefficient of correlation between x and y is found to be 0.72 when determined from a group of 420 boys and 0.65 when determined from a group of 225 girls. Is the difference significant?

Unit - 9

Test of significance (small samples)

When the size of sample is small (<30), the above tests are not applicable. In the large samples, sample S.D. is close to the population S.D. $\therefore$ In the absence of population S.D. we will use sample S.D. But this assumption does not hold good when the sample is small.

When we take samples of size n from a normal population, the variable t defined as $t = \frac{\bar{x} - \mu}{S} \times \sqrt{n}$ where S is the sample S.D. $= \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$. This distribution is known as student - t - distribution. The t -test is not a single distribution but a family of symmetrical distributions distinguished by various values of the parameter v . This parameter is recognised as the degrees of freedom and is equal to the number of observations that can be freely chosen under some overall constraints. Suppose we have 10 values of X which average to $\bar{X}$. Under the constraint that $\bar{x}$ is the same the individual values of x may vary, but nine of them may do so independently. If we chose the values of 9 arbitrary, the 10th is automatically fixed, since $\bar{x}$ is fixed. $\therefore$ With a sample of size n , the variable ' t ' will have a t distribution with $v=(n-1)$ degree of freedom.

9.1 Test of significance based on mean

To test the significant of the mean of a random sample

In determining whether the mean of a sample drawn from a normal population deviates significantly from stated value (the population mean) when variance of the population is unknown,

we calculate the statistic $t = \frac{(\bar{x} - \mu)\sqrt{n}}{s}$ where $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$

Here $\bar{x}$ is the mean of sample.

μ is the mean of the population.

n is the sample size and s - S.D. of the sample.

If the calculated value of $|t| > t_{0.05}$, we say that the difference between $\bar{x}$ and μ is significant at 5% level. If it exceeds $t_{0.01}$, the difference is said to be significant at 1% level.

For 95% the limits are $\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.05}$

For 99% the limits are $\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.01}$

Example: The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with a S.D. of 5 months. A random sample of 6 such bulbs gave the following values.

Life in months 24 26 30 20 20 20 18.

Can you regard the producer's claim to be valid at 1% level of significant?

for $v = 5$, $t_{0.01} = 4.032$.

Solution: Let us take the hypothesis that there is no significant difference in the mean life of bulbs in the sample and that of the population. Applying t test $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$. Let us find s first.

X	$x - \bar{x} (=23)$ (x)	x^2
24	1	1
26	3	9
30	7	49
20	-3	9
20	-3	9
18	-5	25
$\Sigma x = 138$		$\Sigma x^2 = 102$

$$s = \sqrt{\frac{\Sigma x^2}{n-1}} = \sqrt{\frac{102}{5}} = \sqrt{20.4} = 4.517.$$

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{23 - 25}{4.517} \sqrt{6} = \frac{-2}{4.517} \sqrt{6} = -1.084.$$

$$t = 1.084$$

$$v = n - 1 = 6 - 1 = 5.$$

$$t_{0.01} = 4.032.$$

$$t < t_{0.01}.$$

$\therefore$ The hypothesis is accepted.

$\therefore$ The producer's claim is not valid at 1% level.

Example 2: A random sample of size 16 has 53 as mean. The sum of the squares of the deviations taken from mean is 135. Can this sample be regarded as taken from population having 56 as mean? Obtain 95% and 99% confidence limits of the mean of the population.

(For $v = 15$, $t_{0.05} = 2.13$; For $v = 15$, $t_{0.01} = 2.95$)

Solution: Let us take the hypothesis that there is no significant difference between the sample mean and the population mean.

Applying t test,
$$t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

Given $\bar{x} = 53$; $n = 16$; $\mu = 56$.

$$\Sigma (x - \bar{x})^2 = 135.$$

$$\therefore s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} = \sqrt{\frac{135}{15}} = \sqrt{9} = 3.$$

$$\therefore t = \frac{53 - 56}{3} \sqrt{16} = -4.$$

$$|t| = 4.$$

When $v = 15$, $t_{0.05} = 2.13$. The calculated value of t is greater than the table value.

$\therefore$ The hypothesis is rejected.

95% confidence limits of the population mean

$$\begin{aligned} \bar{x} \pm \frac{s}{\sqrt{n}} t_{0.05} &= 53 \pm \frac{3}{\sqrt{16}} \times 2.13 \\ &= 53 \pm 1.6 = 51.4 \text{ to } 54.6. \end{aligned}$$

99% confidence limits are $\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.01}$

$$\begin{aligned} &= 53 \pm \frac{3}{\sqrt{16}} \times 2.95 \\ &= 53 \pm 2.212 = 50.788 \text{ to } 55.212. \end{aligned}$$

Testing difference between means of 2 samples

Independent samples

Given two independent random samples of size n_1 and n_2 with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations s_1 and s_2 we want to test the hypothesis that the samples come from the same normal population. To carry out the test, we calculate the statistic as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad \text{where}$$

$\bar{x}_1$ = means of the first sample

$\bar{x}_2$ = mean of the second sample

n_1 = number of observations in the first sample.

n_2 = number of observations in the second sample

s = combined standard deviation.

The value of s is calculated by the following formula:

$$s = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

When the actual means are in fraction the deviations should be taken from assumed means. In such a case the combined standard deviation is obtained by applying the following formula:

$$s = \sqrt{\frac{\sum (x_1 - A_1)^2 + \sum (x_2 - A_2)^2 - n_1(\bar{x}_1 - A_1)^2 - n_2(\bar{x}_2 - A_2)^2}{n_1 + n_2 - 2}}$$

A_1 = Assumed mean of the first sample

A_2 = Assumed mean of the second sample

$\bar{x}_1$ = Actual mean of the first sample

$\bar{x}_2$ = Actual mean of the second sample

The degrees of freedom = $n_1 + n_2 - 2$

When we are given the number of observations and standard deviation of the two samples, the pooled estimate of standard deviation can be obtained as follows:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

If the calculated value of t be $> t_{0.05}$ ($t_{0.01}$) the difference between the sample means is said to be significant at 5% (1%) level of significance otherwise the data are said to be consistent with the hypothesis.

Example 1: 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 efficacy of the two drugs? If not, which drug should you buy.
 (For $v = 10$, $t_{0.05} = 2.223$)

Solution: Let us take the hypothesis that there is no significant difference in the efficacy of the two drugs. Applying t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
10	-2	4	8	-3	9
12	0	0	9	-2	4
13	1	1	12	1	1
11	-1	1	14	3	9
14	2	4	15	4	16
			10	-1	1
			9	-2	4
$\Sigma x_1 = 60$		$\Sigma (x_1 - \bar{x}_1)^2 = 10$	$\Sigma x_2 = 77$		$\Sigma (x_2 - \bar{x}_2)^2 = 44$

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{60}{5} = 12; \quad \bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{77}{7} = 11.$$

$$s = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{10 + 44}{5 + 7 - 2}} = \sqrt{\frac{54}{10}} = 2.324.$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{12 - 11}{2.324} \sqrt{\frac{5 \times 7}{5 + 7}} = \frac{1.708}{2.324} = 0.735.$$

$$v = n_1 + n_2 - 2 = 5 + 7 - 2 = 10.$$

For $v = 10$, $t_{0.05} = 2.228$.

Since the calculated value of t is less than the table value, the hypothesis is accepted. Hence there is no significance in the efficacy of two drugs. Since drug B is indigenous and there is no difference in the efficacy of imported and indigenous drug, we should buy indigenous drug ie., B.

Example 2: Two laboratories A and B carry out independent estimates of fat content in ice-cream made by a firm. A sample is taken from each batch, halved, and the separate halves sent to the 2 laboratories. The fat content obtained by the labs is recorded below. (The fat content are given in gms).

Batch No:	1	2	3	4	5	6	7	8	9	10
Lab A :	7	8	7	3	8	6	9	4	7	8
Lab B :	9	8	8	4	7	7	9	6	6	6

Is there a significant difference between the mean fat content obtained by the 2 labs A and B. For 18 degrees of freedom, 5% value of t is 2.10.

Solution: Let us take the hypothesis that there is no significant difference between the mean fat content obtained by the two labs A and B.

Applying t- test $t = \frac{\bar{X}_1 - \bar{X}_2}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$

Lab A				Lab B		
Batch No.	X_1	$X_1 - A_1$ $A_1=6$	$(X_1 - A_1)^2$	X_2	$X_2 - A_2$ $A_2=7$	$(X_2 - A_2)^2$
1	7	1	1	9	2	4
2	8	2	4	8	1	1
3	7	1	1	8	1	1
4	3	-3	9	4	-3	9
5	8	2	4	7	0	0
6	6	0	0	7	0	0
7	9	3	9	9	2	4
8	4	-2	4	6	-1	1
9	7	1	1	6	-1	1
10	8	2	4	6	-1	1
	67	7	37	70	0	22

$$\bar{X}_1 = \frac{67}{10} = 6.7; \quad \bar{X}_2 = \frac{70}{10} = 7.$$

Since the actual means are in fractions, we shall take deviations from the assumed means.

$$s = \sqrt{\frac{\sum (x_1 - A_1)^2 + \sum (x_2 - A_2)^2 - n_1(x_1 - A_1)^2 - n_2(x_2 - A_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{37 + 22 - 10(6.7 - 6)^2 - 10(7 - 7)^2}{10 + 10 - 2}} = \sqrt{\frac{59 - 4.9}{19}} = \sqrt{\frac{54.1}{18}} = 1.734$$

$$t = \frac{6.7 - 7}{1.734} \times \sqrt{\frac{10 \times 10}{10 + 10}} = \frac{0.3}{1.734} \times 2.236 = 0.87.$$

For $v = 18$, $t_{0.05} = 2.10$.

The calculated value of $|t|$ is less than the table value. $\therefore$ We accept our hypothesis.

Example 3: The mean life of sample of 10 electric light bulbs was found to be 1,456 hours with S.D. of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1,280 hours with S.D. of 398 hours. Is there a significant difference between the means of the two batches.

Solution: Let us take the hypothesis that the means of the two batches do not differ significantly.

Applying t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Given $\bar{X}_1 = 1456$; $n_1 = 10$; $s_1 = 423$.

$\bar{X}_2 = 1280$; $n_2 = 17$; $s_2 = 398$.

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{9(423)^2 + 16(398)^2}{10 + 17 - 2}}$$

$$= \sqrt{\frac{1610361 + 2534464}{25}} = \sqrt{165793} = 407.18.$$

$$t = \frac{1456 - 1280}{407.18} \times \sqrt{\frac{10 \times 17}{10 + 17}} = \frac{176}{407.18} \times 2.51 = 1.085.$$

$v = 10 + 17 - 2 = 25$.

For $v = 25$, $t_{0.05} = 2.06$.

The calculated value of t is less than the table value. Hence the hypothesis is accepted.

∴ We conclude that the means of the two batches do not differ significantly.

Testing difference between means of 2 samples

Dependent Samples

Two samples are said to be dependent when the elements in one sample are related to those in the other. In fact, the two samples may consist of pairs of observations made on the same selected population elements. Often the use of dependent or paired samples will enable us to perform a more precise analysis, because they will allow us to control the entraneous factors. The t -test based on paired observations is defined by the following formula:

$$t = \frac{\bar{d} - 0}{s} \times \sqrt{n} \quad \text{or} \quad t = \frac{\bar{d}\sqrt{n}}{s}$$

where $\bar{d} = 0$ the mean of the differences

s = the S.D. of the differences

The value of s is calculated as follows:

$$s = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}} \quad \text{or} \quad \sqrt{\frac{\sum d^2 - n(\bar{d})^2}{n - 1}}$$

t is based on $n - 1$ degrees of freedom.

Example 1: To verify whether a course in accounting improved performance, a similar test was given to 12 participants both before and after the course. The original marks - recorded in alphabetical order of the participants were

44 40 61 52 32 44 70 41 67 72 53 72

After the course, the marks were in the same order.

53 38 69 57 46 39 73 48 73 74 60 78

was the course useful.

Solution: Let us take the hypothesis that there is no difference in the marks obtained before and after the course. i.e., the course has not been useful.

Applying t -test, $t = \frac{\bar{d}\sqrt{n}}{s}$

Participants	Before 1 st Test	After 2 nd Test	2 nd - 1 st test d	d ²
A	44	53	9	81
B	40	38	-2	4
C	61	69	8	64
D	52	57	5	25
E	32	46	14	196
F	44	39	-5	25
G	70	73	3	9
H	41	48	7	49
I	67	73	6	36
J	72	74	2	4
K	53	60	7	49
L	72	78	6	36
			$\Sigma d = 60$	$\Sigma d^2 = 578$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{60}{12} = 5.$$

$$s = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{578 - 12(5)^2}{12-1}} = \sqrt{\frac{278}{11}} = 5.03.$$

$$t = \frac{5 \times \sqrt{12}}{5.03} = \frac{5 \times 3.464}{5.03} = 3.443.$$

$$v = n - 1 = 12 - 1 = 11.$$

$$\text{For } v = 11, t_{0.05} = 2.201.$$

The calculated value of t is greater than the table value. So the hypothesis is rejected. Hence the course has been useful.

Example 2: 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?

(5% value of t for 9 degrees of freedom = 2.26).

Solution: Let us take the hypothesis that the drug has no effect on change of blood pressure.

$$\text{Applying t test } t = \frac{\bar{d}\sqrt{n}}{s}$$

d	d ²
3	9
6	36
-2	4
4	16
-3	9
4	16
6	36
6	36
0	0
0	0
2	4
$\Sigma d = 20$	$\Sigma d^2 = 130$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{20}{10} = 2.$$

$$s = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{130 - 10(2)^2}{10-1}} = 3.162.$$

$$t = \frac{2\sqrt{10}}{3.162} = \frac{2 \times 3.162}{3.162} = 2.$$

$$v = n - 1 = 10 - 1 = 9. \text{ For } v = 9, \quad t_{0.05} = 2.26.$$

The calculated value of t is less than the table value. $\therefore$ The hypothesis is accepted. Hence it is reasonable to believe that the drug has no effect on change of blood pressure.

Testing the significance of an observed correlation coefficient

Given a random sample from a bivariate normal population. If we are to test the hypothesis that the correlation coefficient of the population is zero, i.e., the variables in the population are uncorrelated, we have to apply the following test:

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$$

Here t is based on $(n - 2)$ degrees of freedom.

If the calculated value of t exceeds $t_{0.05}$ for $(n-2)$ d.f. (degrees of freedom) we say that the value of r is significant at 5% level. If $t < t_{0.05}$, the data are consistent with the hypothesis of an uncorrelated population.

Example 1: A study of the heights of 18 pairs of husbands and their wives in a factory shows that the coefficient correlation is 0.52. Apply t - test to find whether correlation is significant. (For 16 d.f. at 5% level of significance, the table value of $t = 2.12$).

Solution: Let us take the hypothesis that there is no significant difference in the sample correlation and correlation in the population. Applying t - test:

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

Given $r = 0.52$; $n = 18$.

$$t = \frac{0.52}{\sqrt{1-(0.52)^2}} \times \sqrt{18-2} = \frac{0.52 \times 4}{0.854} = 2.44$$

$$v = n - 2 = 18 - 2 = 16.$$

For $v = 16$, $t_{0.05} = 2.12$.

The calculated value of t is greater than the table value. So the hypothesis is rejected. $\therefore$ The given value of r is significant.

Example 2: Find the least value of r in a sample of 27 pairs from a bivariate normal population significant at 5% level.

Solution: $t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$

$$= \frac{r}{\sqrt{1-r^2}} \times \sqrt{27-2} = \frac{5r}{\sqrt{1-r^2}}$$

For ' r ' to be significant at 5% level, calculated value of t should be greater than the table value.

At 5% level for 25 d.f. the table value is 2.06.

$$\therefore \frac{5r}{\sqrt{1-r^2}} > 2.06$$

$$25r^2 > 4.2436(1-r^2).$$

$$29.2436r^2 > 4.2436.$$

$$r > \sqrt{\frac{4.2436}{29.2436}} = 0.381.$$

Hence the required least value of r in a sample of 27 pairs to be significant at 5% level should be 0.381.

Exercise

1. The life time of electric bulbs for a random sample of 10 from a large consignment gave the following data:

Item	1	2	3	4	5	6	7	8	9	10
Life in 1000 hours:	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the average life time of bulbs is 4,000 hours.

2. For a random sample of 10 persons, fed on diet A the increase to weight in pounds in a certain period were:

10, 6, 16, 17, 13, 12, 8, 14, 15, 9.

For another random sample of 12 persons, fed on diet B, the increase in the same period were

7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17.

Test whether the diets A and B differ significantly as regards their effect on increase in weight.

For $v = 20$, $t_{0.05} = 2.09$.

3. In a test given to two groups of students, the marks obtained are as follows:

First group	:	18	20	36	50	49	36	34	49	41
Second group	:	29	28	26	35	30	44	46		

Examine the significance of difference between the arithmetic mean of the marks secured by the students of the above two groups.

(The value of t at 5% level of significance for $v = 14$ is 2.14).

4. An I. Q. test was administered to 5 persons before and after they were trained. The results are given below.

		I	II	III	IV	V
I.Q. before training	:	110	120	123	132	125
I.Q. after training	:	120	118	125	136	121

Test whether there is any change in I.Q. after the training programme. [For $v = 4$, $t_{0.01} = 4.6$].

5. Is a correlation coefficient of 0.5 significant if obtained from a random sample of 11 pairs of values from a normal population? Use t -test.

6. A group of 5 patients treated with medicine A weigh 42, 39, 48, 60, 41 kgs. Second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69, 62 kgs. Do you agree with the claim that medicine B increases the weight significantly.

9.2 Test of significance based on F- test

Here we see test of significance for equality of population variances using Snedecor's F-distribution. Suppose we want to test whether two independent samples $x_1, x_2, \dots, x_{n_1}$ and $y_1, y_2, \dots, y_{n_2}$ have been drawn from the normal population with the same variance σ^2 .

$$\text{Define } F = \frac{u^2}{v^2} \text{ where } u^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n_1 - 1} = \frac{n_1 s_1^2}{n_1 - 1}$$

$$\text{and } v^2 = \frac{\sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_2 - 1} = \frac{n_2 s_2^2}{n_2 - 1}$$

It follows Snedecor's F- distribution with (n_1-1, n_2-1) degrees of freedom considering the greater of u^2 and v^2 always in the numerator of F.

Set the null hypothesis $H_0 : \sigma_x^2 = \sigma_y^2 = \sigma^2$.

Under the null hypothesis $F = \frac{n_1 s_1^2 / (n_1 - 1)}{n_2 s_2^2 / (n_2 - 1)}$ is a test statistic.

By comparing the calculated value of F for the two given samples with the tabulated value of F for $(n_1 - 1, n_2 - 1)$ d.f. at any required level of significance H_0 is either accepted or rejected.

Example 1: Test the equality of s.d.s for the data given below at 5% level of significance $n_1=10$; $n_2=14$; $s_1=1.5$; $s_2=1.2$.

Solution: Set the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$.

$$\begin{aligned} \text{Under } H_0, F &= \frac{n_1 s_1^2 (n_1 - 1)}{n_2 s_2^2 (n_2 - 1)} = \frac{(10 \times (1.5)^2) / 9}{(14 \times (1.2)^2) / 13} \\ &= \frac{(10 \times 2.25) / 9}{(14 \times 1.44) / 13} = \frac{22.5 / 9}{20.16 / 13} = \frac{2.5}{1.55} = 1.61. \end{aligned}$$

The table value for (9, 13) d.f. at 5% point level of significance is 2.72. Calculated $F = 1.61 < 2.72$.
 $\therefore$ It is not significant. Hence H_0 may be accepted.

Example 2: in one sample of 8 observations the sum of the squares of deviation of the sample values from the sample mean was 84.4 and in another sample of 10 observations it was 102.6. Test whether the difference in variances is significant at 5% level using F- test.

Solution: Given $n_1 = 8$; $n_2 = 10$.

$$\sum_{i=1}^8 (x_i - \bar{x})^2 = 84.4.$$

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 102.6$$

Set the null hypothesis $H_0 : \sigma_1 = \sigma_2$.

Under this null hypothesis $F = \frac{n_1 s_1^2 / (n_1 - 1)}{n_2 s_2^2 / (n_2 - 1)}$

where $s_1^2 = \frac{\sum (x_i - \bar{x})^2}{8}$ and $s_2^2 = \frac{\sum (y_i - \bar{y})^2}{10}$

$$\therefore F = \frac{84.4 / 7}{102.6 / 9} = \frac{12.057}{11.42} = 1.056.$$

The table value for (7,9) d.f. at 5% level of significance is 3.29.

$F = 1.056 < 3.29$. $\therefore$ The difference is not significant.

Example 3: Two random samples drawn from 2 normal populations are given below. Test whether the 2 populations have the same variance.

Sample I	20	16	26	27	23	22	18	24	25	19		n=10	
Sample II	17	23	32	25	22	24	28	6	31	33	30	27	n=12

Solution: Given $n_1 = 10$; $n_2 = 12$.

Mean of the first sample $\bar{x} = \frac{20 + 16 + \dots + 19}{10} = \frac{220}{10} = 22$.

Mean of the second sample $\bar{y} = \frac{288}{12} = 24$.

$$s_1^2 = \frac{\sum (x_i - 22)^2}{10} = \frac{2^2 + 6^2 + 4^2 + 5^2 + 1 + 0 + 4^2 + 2^2 + 3^2 + 3^2}{10}$$

$$\therefore 10s_1^2 = 120.$$

Similarly $s_2^2 = \frac{\sum (y_i - 25)^2}{12} = \frac{614}{12}$

$$\therefore 12s_2^2 = 614.$$

Set the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$.

$$n_1 s_1^2 / (n_1 - 1) = 10s_1^2 / 9 = 120 / 9 = 13.33.$$

$$n_2 s_2^2 / (n_2 - 1) = 12s_2^2 / 11 = 614 / 11 = 55.82.$$

Under the null hypothesis the test statistic

$$F = \frac{n_1 s_1^2 / n_1 - 1}{n_2 s_2^2 / n_2 - 1} = \frac{55.82}{13.33} = 4.19 \text{ (approximately)}$$

The table value for (11, 9) d.f. at 5% level of significance is 3.10. $F = 4.19 > 3.10$. $\therefore$ The difference is significant. Hence we reject the hypothesis. i.e., the 2 populations do not have the same variance.

Exercise

1. For 2 samples of sizes 8 and 12 the observed variances are 0.064 and 0.024. Test the hypothesis that the samples came from normal populations with variances equal.
2. In a sample of 8 observations the sum of the squared deviations of items from the mean was 94.5. In another sample of 10 observations the value was found to be 101.7. Test whether the difference is significant.
3. Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means are equal to 160 and 91 respectively. Can they be regarded as drawn from the same normal populations?
4. The nicotine content in milligrams of two samples of tobacco were found to be as follows. Can it be said that the 2 samples come from the same normal population.

Sample A	24	27	26	21	25	-
Sample B	27	30	28	31	22	36

9.3 Test for significance of an observed sample correlation

Given a bivariate normal population. Consider a random sample $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ from the above population. Let r be the correlation coefficient of the random sample. When the variables in the populations are uncorrelated i.e., $\rho = 0$ it is found that the variable

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} \text{ follows student's } t\text{-distribution with } (n-2) \text{ d.f.}$$

Setting the null hypothesis $H_0 : \rho = 0$, $t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$ becomes a test statistic. It can be tested at any required level of significance to test whether the sample could have come from the uncorrelated population or not.

Example 1: For a sample of size 19 a correlation of 0.36 was obtained. Can this be from a population with correlation coefficient $\rho = 0$.

Solution: Given $n=19$; $r=0.36$. d.f. $v=17$.

Set the null hypothesis $H_0 : \rho = 0$.

The test statistic is

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} = \frac{0.36\sqrt{17}}{\sqrt{1-(0.36)^2}} = \frac{0.36 \times 4.123}{\sqrt{0.8704}}$$

$$= \frac{1.484}{0.932} = 1.592.$$

The table value for 17 d.f. at 5% level of significance is 2.11.

$$\therefore |t| = 1.592 < 2.11 = t_{0.05}.$$

$\therefore$ It is not significant.

Hence H_0 may be accepted at 5% level of significance.

$\therefore$ The sample could have come from an uncorrelated population.

Example 2: Find the least value of r in a sample of 11 pairs from a bivariate normal population significant at (i) 5% level (ii) 1% level.

Solution: Given $n=11$.

(i) $|t| > t_{0.05}$ for 9 d.f.

(ii) $|t| > t_{0.01}$ for 9 d.f. where $t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$.

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{9} = \frac{3r}{\sqrt{1-r^2}}$$

The table value of $t_{0.05}$ for 9 d.f. is 2.26 and $t_{0.01}$ for 9 d.f. is 3.25.

$$(i) |t| > 2.26 \Rightarrow \left| \frac{3r}{\sqrt{1-r^2}} \right| > 2.26 \rightarrow \frac{9r^2}{1-r^2} > 5.1076$$

$$\therefore 14.1076r^2 > 5.1076.$$

$$\therefore r^2 > \frac{5.1076}{14.1076} = 0.36.$$

$\therefore |r| > 0.6$. Hence the least value of r significant at 5% level is 0.6.

$$(ii) |t| > 3.25 \Rightarrow \left| \frac{3r}{\sqrt{1-r^2}} \right| > 3.25.$$

$$\Rightarrow \frac{9r^2}{1-r^2} > 10.5625.$$

$$9r^2 > (10.5625)(1-r^2)$$

$$|r| > 0.73.$$

$\therefore$ The least value of r significant at 1% level is 0.73.

Exercise

1. A random sample of 10 observations gave a correlation of 0.2. Is there significant of correlation in the population?
2. Find the least value of r in a sample of 27 pairs from a bivariate normal population significant at 5% level.
3. A random sample of 18 pairs from a bivariate normal population showed a correlation coefficient of 0.3. Does this show evidence of having come from a population with zero correlation?
4. Test the significance of the following rank correlation coefficient (i) $r = 0.139$; (ii) $r = 0.618$; $n = 20$.

Unit 10

Test based on χ^2 - Distribution

The χ^2 - distribution has number of applications in statistics. Here we deal with three important applications based on χ^2 - distribution.

- I χ^2 - test for population variance.
- ii χ^2 - test to test the goodness of fit.
- III χ^2 - test to test the independence of attributes.

10.1 χ^2 - test for population variance

When we want to test if a random sample of $x_1, x_2, \dots, x_n$ has been drawn from a normal population with mean μ and variance σ^2 , the statistic $\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\sigma^2} = \frac{ns^2}{\sigma^2}$ where s , the S.D. of the sample follows χ^2 - distribution with $(n-1)$ degrees of freedom. By comparing the calculated value of χ^2 with the tabulated value $(n-1)$ d.f. at certain level of significant we may accept or reject the null hypothesis.

It should be noted that this test can be applied only if the population is normal.

Example 1: A random sample of size 25 from a population gives the sample S.D. to be 8.5. Test the hypothesis that the population S.D. is 10.

Solution: We set the null hypothesis H_0 that the population S.D. is 10. ie. $\sigma = 10$.

Given $n = 25$, $s = 8.5$.

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{25(8.5)^2}{10^2} = \frac{1806.25}{100} = 18.0625.$$

$$v = n - 1 = 25 - 1 = 24.$$

For $v=24$, $\chi^2_{0.05} = 36.145$

Since the calculated value of χ^2 is less than the table value we accept our hypothesis.

Example 2: Weights in kg of 10 students are given below: 38, 40, 45, 53, 47, 43, 55, 48, 52, 49. Can we say that variance of the distribution of weights of all students from which the above sample of 20 students was drawn is equal to 20 square kg? For 9 d.f. $\chi^2_{0.05} = 16.92$.

Solution: Applying χ^2 test:

x	$x - \bar{x}$	$(x - \bar{x})^2$
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	16
49	+ 2	4
$\Sigma x = 470$	$\Sigma (x - \bar{x}) = 0$	$\Sigma (x - \bar{x})^2 = 280$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{470}{10} = 47.$$

$$\chi^2 = \frac{\Sigma (x_i - \bar{x})^2}{\sigma^2} = \frac{280}{20} = 14.$$

$$v = 10 - 1 = 9.$$

$$\text{For } v = 9, \quad \chi^2_{0.05} = 16.92.$$

The calculated value of χ^2 is less than the table value. $\therefore$ We conclude that the population variance may be 20 square kg.

The χ^2 test when the degrees of freedom exceed 30.

The table values of χ^2 are available only upto 30 degrees of freedom. For greater than 30, the distribution of $\sqrt{2}\chi^2$ approximates the normal distribution. The mean of the distribution $\sqrt{2}\chi^2$ is $\sqrt{2n-1}$ and the S.D. is = 1.

$$\therefore z = \frac{\sqrt{2}\chi^2 - \sqrt{2n-1}}{1}.$$

Example: Test the hypothesis that $\sigma = 8$ given that $s = 10$ for a random sample of size 51.

Solution: Given $n=51$; $\sigma = 8$; $s = 10$

$$\text{Let } H_0 : \sigma = 8$$

Under the null hypothesis the test statistic is

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{51 \times 10^2}{8^2} = \frac{5100}{64} = 79.7.$$

Since $n = 51 > 30$, we use the test statistics the standard normal variate $z = \sqrt{2\chi^2} - \sqrt{2n-1}$

$$\begin{aligned} z &= \sqrt{2 \times 79.7} - \sqrt{2 \times 51 - 1} \\ &= 12.63 - 10.05 = 2.58. \end{aligned}$$

$|z| = 2.58 > 1.96$. Hence the difference is significant at 5% level of significance and hence the hypothesis is rejected at 5% level of significance.

Exercise

1. A random sample of size 20 from a population gives the sample S.D. of 6. Test the hypothesis that the population S.D. is 9.
2. A sample of 12 values shows the S.D. to be 11. Does this agree with the hypothesis that the population S.D. is 10, the population being normal?
3. Test the hypothesis that $\sigma = 10$ given that $s = 15$ for a random sample of size 50 from a normal population.

10.2 - χ^2 - test to test the goodness of fit

This test can be applied to test for compatibility of observed frequencies and theoretical frequencies. Let $O_1, O_2, \dots, O_n$ be the observed frequencies and $e_1, e_2, \dots, e_n$ be the corresponding expected frequencies such that $\sum_{i=1}^n O_i = N = \sum_{i=1}^n e_i$ where N is the number of members in the population.

Define $\chi^2 = \sum_{i=1}^n \frac{(O_i - e_i)^2}{e_i}$ It is a χ^2 variable with $n - 1$ degrees of freedom. Setting the null

hypothesis H_0 : the observed and theoretical frequencies are compatible, the value of the test statistic χ^2 is found out.

By comparing the calculated value with the tabulated value of χ^2 for $n - 1$ degrees of freedom at any required level of significance we may accept or reject the null hypothesis.

Note 1: Both experiments and theoretical considerations show that χ^2 - approximation will be satisfactory if all $e_i \geq 5$ and $n \geq 5$. If $n < 5$, it is better to have e_i somewhat larger than 5. If some of the frequencies O_i do not exceed 5 they may be combined with other frequencies so that the condition is satisfied. Correspondingly the degrees of freedom will be reduced.

Note 2: Suppose m parameters of the population are not specified but have to be estimated from the data itself. This imposes m further constraints to be satisfied by O_i^s and e_i^s . If all the constraints imposed in this process are linear then the number of degrees of freedom is $n-1-m$.

Example 1: The following table gives the number of train accidents in a country that occurred during the various days of the week. Find whether the accidents are uniformly distributed over the week.

Days	Sun	Mon	Tue	Wed	Thur	Fri	Sat
No. of accidents	20	18	13	23	26	11	15

Solution: Given $n = 7$.

Total number of the accidents in the week = 126.

If the accidents are to be uniformly distributed it is expected $\frac{126}{7} = 18$ accidents to happen per day. Hence $e_i = 18$ for all the days of the week. The observed and expected number of accidents are given below.

O_i	20	18	13	23	26	11	15
e_i	18	18	18	18	18	18	18

Set up the null hypothesis H_0 : Accidents are uniformly distributed.

$$\begin{aligned}\chi^2 &= \frac{\sum (O_i - e_i)^2}{e_i} \\ &= \frac{4}{18} + 0 + \frac{25}{18} + \frac{25}{18} + \frac{64}{18} + \frac{49}{18} + \frac{9}{18} \\ &= \frac{176}{18} = 9.78.\end{aligned}$$

Degrees of freedom $n - 1 = 6$.

The table value of χ^2 for 6 d.f. is 12.592 at 5% level of significance.

The calculated value of $\chi^2 = 9.78 < 12.592$.

$\therefore$ It is not significant at 5% level of significance. So we accept the hypothesis and hence we conclude that the accidents are uniformly distributed over the week.

Example 2: Five coins are tossed 320 times. The number of heads observed is given below. Examine whether the coin is unbiased.

No. of heads	0	1	2	3	4	5	Total
Frequency	15	45	85	95	60	20	320

Solution: The chance of getting a head in a single throw $p = \frac{1}{2}$. $\therefore q = \frac{1}{2}$.

Given $n=5$ and $N = 320$.

The expected frequencies $e(x)$ for x heads in N trials when n coins are tossed are given by $N n C_x p^x q^{n-x}$.

When 5 coins are tossed, the expected frequencies for 0, 1, 2, 3, 4, 5 heads in 320 trials are given

$$\text{by } e(0) = 320 \times 5 C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = \frac{320}{32} = 10.$$

$$e(1) = 320 \times 5 C_1 \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^4 = \frac{1600}{32} = 50.$$

$$e(2) = 320 \times 5 C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = \frac{3200}{32} = 100.$$

$$e(3) = 320 \times 5 C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = \frac{3200}{32} = 100.$$

$$e(4) = 320 \times 5 C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right) = \frac{1600}{32} = 50.$$

$$e(5) = 320 \times \left(\frac{1}{2}\right)^5 = \frac{320}{32} = 10.$$

The observed and expected values are given below.

O_i	15	45	85	95	60	20
e_i	10	50	100	100	50	10

Set up the null hypothesis H_0 : The coin is unbiased. ie., $p = \frac{1}{2}$.

Under the null hypothesis the test statistic is

$$\chi^2 = \frac{\sum (O_i - e_i)^2}{e_i}$$

$$= \frac{5^2}{10} + \frac{5^2}{50} + \frac{15^2}{100} + \frac{5^2}{100} + \frac{10^2}{50} + \frac{10^2}{10} = 17.5$$

The d.f. is $6 - 1 = 5$. The table value of χ^2 for 5 d.f. at 5% level of significance is 11.07. Calculated value of $\chi^2 = 17.5 > 11.07$.

It is significant at 5% level of significance. Hence the hypothesis may be rejected at 5% level of significance. So the coin is biased.

Example 3: Fit a Poisson distribution for the following data and test the goodness of fit.

x	0	1	2	3	4	5	6	Total
f	273	70	30	7	7	2	1	390

Solution: $\bar{x} = \frac{\sum fx}{\sum f} = \frac{70 + 60 + 21 + 28 + 10 + 6}{390} = \frac{195}{390} = \frac{1}{2}$

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

The expected frequencies of the Poisson distribution are given by $f(x) = \frac{N e^{-\lambda} \lambda^x}{x!} = \frac{390 \left(\frac{1}{2}\right)^x}{e x!}$
 $x = 0, 1, 2, 3, 4, 5, 6$

$$f(0) = \frac{390}{e} \left(\frac{1}{2}\right)^0 = 236.4$$

$$f(1) = \frac{390}{e 1!} \left(\frac{1}{2}\right) = \frac{390}{1.65} \left(\frac{1}{2}\right) = 118.2$$

$$f(2) = \frac{390}{e 2!} \left(\frac{1}{2}\right)^2 = 29.50$$

$$f(3) = \frac{390}{e 3!} \left(\frac{1}{2}\right)^3 = 4.9$$

$$f(4) = \frac{390}{e 4!} \left(\frac{1}{2}\right)^4 = 0.6$$

$$f(5) = \frac{390}{e 5!} \left(\frac{1}{2}\right)^5 = 0.1$$

$$f(6) = \frac{390}{e 5!} \left(\frac{1}{2}\right)^6 = 0.005.$$

$\therefore$ The observed and expected frequencies are:

O _i	273	70	30	7	7	2	1	390
e _i	236.4	118.2	29.5	4.9	0.6	0.1	0	389.7

Since the sum of the expected frequencies is 389.7 which is 0.3 less than 390, it can be adjusted in the last two frequencies by adding 0.3. Since the frequencies in the last three cell frequencies are less than 5 they can be pooled as one cell and we get the regrouped O_i and e_i as

O_i	273	70	30	17	390
e_i	236.4	118.2	29.5	5.9	390

Set up the null hypothesis H_0 : Poisson distribution can be fitted well.

$$\chi^2 = \frac{\sum (O_i - e_i)^2}{e_i}$$

$$= \frac{(273 - 236.4)^2}{236.4} + \frac{(70 - 118.2)^2}{118.2} + \frac{(30 - 29.5)^2}{29.5} + \frac{(17 - 5.9)^2}{5.9} = 46.3.$$

There are 2 linear constraints used one for the total and another for calculating $\bar{x}$ from the observed value.

$$\therefore \text{Degrees of freedom} = 7 - 1 - 1 - 3 = 2.$$

The table value of 2 d.f. at 5% level is 5.99. Since calculated $\chi^2 = 46.3 > 5.99$, it is much significant at 5% level of significance.

$\therefore$ The hypothesis is rejected at 5% level and we may conclude that Poisson distribution is not a good fit to the data.

Exercise:

1. In 200 toss of a coin 115 head and 85 tails were observed. Test the hypothesis whether the coin is unbiased at 5% level of significance using χ^2 .
2. The following table gives the number of aircraft accidents that occurred during the seven days of a week. Find whether the accidents are uniformly distributed over the week.

Days	S	M	T	W	Th	F	Sat	Total
No. of accidents	14	16	8	12	11	9	14	84

3. In a set of random numbers the digits 0, 1, 2, ..., 9 were found to have the following frequencies. Test whether they are significantly different from those expected on the hypothesis of uniform distributions.

Digits	0	1	2	3	4	5	6	7	8	9
Frequencies	43	32	38	27	38	52	36	31	39	24

4. In 120 throws of a single die the following distribution of faces was obtained. Do these data indicate an unbiased die?

Faces	1	2	3	4	5	6	Total
Frequencies	30	25	18	10	22	15	120

5. When the first proof of 200 pages of a book of 5000 pages were read the distribution of printing mistakes were found to be follows. Fit a Poisson distribution to the above data and test the goodness of fit.

No. of mistakes in a page	0	1	2	3	4	5
No. of pages	12	63	20	3	1	1

6. A survey of 320 families with 5 children each revealed the following distribution. Is this result consistent with the hypothesis that male and female births are equally probable?

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

7. The following data represents the monthly sales in rupees of a certain retail shop in a year. Examine if there is any seasonality in the sales.
610, 560, 635, 605, 625, 620, 630, 625, 580, 600, 615, 615

8. The following table shows the observed and expected frequencies in tossing a die 120 times. Test whether the die is fair at 5% level of significance.

Face	1	2	3	4	5	6
O_i	25	17	15	23	24	16
e_i	20	20	20	20	20	20

10.3 – χ^2 test to test the independence of attributes

With the help of χ^2 test we can find out whether two or more attributes are associated or not. Suppose we have N observations classified according to some attributes. We may ask whether the attributes are related or independent. Thus, we can find out whether quinine is effective in controlling fever or not, whether there is any association between marriage and failure, or eye colour of husband and wife. In order to test whether or not the attributes are associated we take the null hypothesis that there is no association in the attributes under study.

If the calculated value of χ^2 for given d.f. is less than the table value at a certain level of significance we say that the attributes are not associated. If the calculated value of χ^2 for given d.f. is greater than the table value at a certain level of significance, we say that the attributes are associated. It should be noted that χ^2 is not a measure of the degree or form of relationship, it only tells us whether two principles of classification are or are not significantly related without reference to any assumption concerning the form of relationship.

In a contingency table the degrees of freedom are calculated thus follows:

The marginal total or frequencies place the limit on our choice of selecting cell frequencies. The cell frequencies of all columns but one (c-1) and of all rows but one (r-1) can be assigned arbitrarily and so the number of d.f. for all the cell frequencies = (c-1) (r-1) where c refers to columns and r refers to rows. Thus in a 2x2 table the d.f. = (2-1) (2-1) = 1. Similarly in a 3x3 contingency table, the number of d.f. is (3-1) (3-1) = 4 and so on.

To calculate the expected frequencies we have $E = \frac{RT \times CT}{N}$ where

E - Expected frequency

RT - Row total for the row containing the cell

CT - Column total for the column containing the cell.

N - The total number of observations.

The value of χ^2 is given by $\chi^2 = \frac{\sum (O - E)^2}{E}$ where O is the observed frequency.

The value of χ^2 can range from zero to infinity. If χ^2 is zero it means that the observed and expected frequencies completely coincide.

If any one of the observed frequencies is less than 10, we will use Yate's correction and then apply χ^2 test.

$$\chi^2 \text{ corrected} = \frac{\sum (|O - E| - 0.5)^2}{E} \text{ or reduce 0.5 for 2 cells and increase 0.5 for 2 cells.}$$

This correction is made only when the d.f. is 1 and N is small.

Example 1: In an experiment on immunization of cattle from tuberculosis the following results are obtained.

	Affected (B)	Not affected (β)
Inoculated (A)	12	26
Not inoculated (α)	16	6

Calculate χ^2 and discuss the effect of vaccine in controlling susceptibility to tuberculosis. (5% value of χ^2 for 1 d.f. = 3.84).

Solution: Let us take the hypothesis that the vaccine is not effective in controlling susceptibility to tuberculosis.

Before applying χ^2 test we have find expected frequencies.

$$E \text{ of } (A\beta) = \frac{(12+26)(26+6)}{60} = \frac{38 \times 32}{60} = 20.3$$

$$E \text{ of } (AB) = \frac{(12 + 26) \times (12 + 16)}{60} = 17.7.$$

$$E \text{ of } (\alpha B) = \frac{(12 + 26) \times (16 + 6)}{60} = \frac{28 \times 22}{60} = 10.3.$$

$$E \text{ of } (\alpha \beta) = \frac{(16 + 6) \times (26 + 6)}{60} = \frac{22 \times 32}{60} = 11.7.$$

The expected frequencies are

17.7	20.3	38
10.3	11.7	22
28	32	60

Since one of the observed frequencies is less than 10, we will use Yates' correction and then apply χ^2 test.

O	E	(O - E) ²	(O - E) ² / E
12.5	17.7	27.04	1.528
15.5	10.3	27.04	2.625
25.5	20.3	27.04	1.332
6.5	11.7	27.04	2.311
			7.796

$$\Sigma (O - E) = \Sigma O - \Sigma E = N - N = 0.$$

$$\chi^2 = 7.796. \quad v = 1.$$

$$\text{For } v = 1, \chi^2_{0.05} = 3.84.$$

Since the calculated value of χ^2 is greater than the table value we reject our hypothesis.

Example 2: A certain drug is claimed to be effective in curing colds. In an experiment on 328 people with colds, half of them were given the drug and half of them given sugar pills. The patients reactions to the treatment are recorded in the following table. Test the hypothesis that the drug is no better than sugar pills for curing colds (Given for $v=2$, $\chi^2_{0.05} = 5.99$).

	Helped	Harmed	No effect
Drug	104	20	40
Sugar Pills	88	24	52

Solution: Let us take the hypothesis that there is no significant difference in the effect of the drug and sugar pills. The given frequencies are

104	20	40	164
88	24	52	164
192	44	92	328

$$E_{11} = \frac{164 \times 192}{328} = 96$$

$$E_{21} = \frac{164 \times 192}{328} = 96$$

$$E_{12} = \frac{164 \times 44}{328} = 22$$

$$E_{22} = \frac{164 \times 44}{328} = 22$$

$$E_{13} = \frac{164 \times 92}{328} = 46$$

$$E_{23} = \frac{164 \times 92}{328} = 46$$

The expected frequencies are

96	22	46	164
96	22	46	164
192	44	92	328

Applying χ^2 test:

O	E	(O-E) ²	(O-E) ² / E
104	96	64	0.667
88	96	64	0.667
20	22	4	0.182
24	22	4	0.182
40	46	36	0.783
52	46	36	0.783
$\Sigma[(O-E)^2/E] = 3.264$			

$$\chi^2 = \Sigma(O-E)^2/E = 3.264$$

$$v = (r-1)(c-1) = (2-1)(3-1) = 2.$$

$$\text{For } v=2, \chi^2_{0.05} = 5.991.$$

The calculated value of χ^2 is less than the table value and hence the result of the experiment shows that there is no significant difference in the effect of the drug and sugar pills.

Example 3: A movie producer is bringing out a new movie. In order to map out his advertising campaign he wants to determine whether the movie will appeal most to a particular age group or whether it will appeal equally to all age groups, the producer takes a random sample from persons attending preview of the new movie, and obtains the following results:

	Age groups				
	Under 20	20-39	40-59	60 & over	Total
Liked the movie	146	78	48	28	300
Disliked the movie	54	22	42	22	140
Indifferent	20	10	10	20	60
Total	220	110	100	70	500

What inference will you draw from this data?

Solution: Let us take the hypothesis that the movie appeals equally to all age groups. The expected frequencies are:

$$E_{11} = \frac{300 \times 220}{500} = 132$$

$$E_{21} = \frac{140 \times 220}{500} = 61.6$$

$$E_{12} = \frac{300 \times 110}{500} = 66$$

$$E_{22} = \frac{140 \times 110}{500} = 30.8$$

$$E_{13} = \frac{300 \times 100}{500} = 60$$

$$E_{23} = \frac{140 \times 100}{500} = 28$$

$$E_{14} = \frac{300 \times 70}{500} = 42$$

$$E_{24} = \frac{140 \times 70}{500} = 19.6$$

$$E_{31} = \frac{60 \times 220}{500} = 26.4$$

$$E_{32} = \frac{60 \times 110}{500} = 13.2$$

$$E_{33} = \frac{60 \times 100}{500} = 12$$

$$E_{34} = \frac{60 \times 100}{500} = 8.4$$

The tabulated values are:

132.0	66.0	60	42	300
61.6	30.8	28	19.6	140
26.4	13.2	12	8.4	60
220	110	100	70	500

O	E	(O-E) ²	(O-E) ² / E
146	132	196	1.485
54	61.6	57.76	0.938
20	26.4	40.90	1.552
78	66	144	2.182
22	30.8	77.44	2.154
10	13.2	10.24	0.776
48	60	144	2.400
42	28	196	7.0
10	12	4	0.333
28	42	196	4.667
22	19.6	5.76	0.294
20	8.4	134.56	16.019
			40.16

Applying χ^2 test

$$\chi^2 = \sum (O-E)^2 / E = 40.16$$

$$v = (r-1) (c-1) = (3-1) (4-1) = 6$$

$$\text{For } v = 6, \chi^2_{0.05} = 12.59.$$

The calculated value of χ^2 is much greater than the table value. So, the hypothesis is rejected.
 $\therefore$ we conclude that the movie does not appeal equally to all age groups.

Exercise

- 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	I.Q.		Total
	High	Low	
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

Given for $v = 1$, $\chi^2_{0.05} = 3.84$

- An automobile company gives you the following information about age groups and the liking for particular model of car which it plans to introduce.

Persons who:	Age group				Total
	Below 20	20-39	40-59	60 & above	
Liked the car	140	80	40	20	280
Disliked the car	60	50	30	80	220
Total	200	130	70	100	500

On the basis of this data can it be concluded that the model appeal is independent of the age group (Given for $v=3$, $\chi^2_{0.05} = 7.815$).

- The contingency table below summarizes the results obtained in a study conducted by a research organisation, with respect to the performance of four competing brands of toothpaste among the users:

	Brand A	Brand B	Brand C	Brand D	Total
No cavities	9	13	17	11	50
One to five cavities	63	70	85	82	300
More than five cavities	28	37	48	37	150
Total	100	120	150	130	500

Test the hypothesis that incidence of cavities is independent of the brand of the toothpaste used.

(For $v=6$, $\chi^2_{0.05} = 12.59$).

4. From the following data find out whether there is any relationship between sex and preference of colour:

Colour	Males	Females	Total
Red	10	40	50
White	70	30	100
Green	30	20	50
Total	110	90	200

5. A survey amongst women was conducted to study the family life. The observations are as follows:

	Family Life		Total
	Happy	Not Happy	
Educated	70	30	100
Not educated	60	40	100
Total	130	70	200

Test whether there is any association between family life and education.

10.4- Analysis of variance

The analysis of variance frequently referred to by the contraction ANOVA is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal.

The analysis of variance technique developed by R.A. Fisher in 1920's is capable of fruitful application to a diversity of practical problems. Basically, it consists of classifying and cross-classifying statistical results and testing whether the means of a specified classification differ significantly. In this way it is determined whether the given classification is important in affecting the results.

For example, five fertilizers are applied to four plots each of wheat and we are given the yield of wheat on each of these plots. We may be interested in finding out whether the effects of these fertilizers on the yields are significantly different or whether the samples have come from the same universe. The answer to this problem is provided by the technique of analysis of variance.

Technique of analysis of variance

For the sake of clarity the technique of analysis of variance has been discussed separately for (a) one way classification and (b) two-way classification.

One way classification

In one way classification the data are classified according to only one criterion.

The null hypothesis is:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \neq \mu_k$. All the means are not equal. The steps in carrying out the analysis are:

1) Calculate variance between the samples

The steps in calculating variance between samples will be:

a) Calculate the mean of each sample i.e., $\bar{x}_1, \bar{x}_2$, etc;

b) Calculate the grand average $\bar{\bar{x}}$. It's value is obtained by
$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \dots}{N_1 + N_2 + N_3 + \dots}$$

c) Take the difference between the means of the various sample and the grand average.

d) Square these deviations and obtain the total which will give sum of squares between the samples ; and

e) Divide the total obtained in step (d) by the degrees of freedom. The degrees of freedom will be one less than the number of samples.

2) Calculate variance within the samples

The steps in calculating variance within the samples will be:

a) Calculate the mean value of each sample i.e., $\bar{x}_1, \bar{x}_2, \bar{x}_3$, etc.,

- b) Take the deviations of the various items in a sample from the mean values of the respective samples.
- c) Square these deviations and obtain the total which gives the sum of squares within the samples and
- d) Divide the total variance in step (c) by the degrees of freedom. The degree of freedom is given by $v=N-k$ where k refers to the number of samples and N refers to the total number of all the observations.

3) Calculate the ratio **F** as follows:

$$F = \frac{\text{Between column variance}}{\text{Within column variance}}$$

4) Compare the calculated value of **F** with the table value of **F** for the degrees of freedom at a certain critical level. If the calculated value of **F** is greater than the table value, the difference in sample means is significant. If the calculated value of **F** is less than the table value the difference is not significant.

Analysis of Variance (ANOVA) Table : One-way classification Model

Source of Variation	SS (Sum of squares)	V Degrees of freedom	MS Mean square	Variance Ratio of F
Between Samples	SSC	$v_1=c-1$	$MSC=SSC/(c-1)$	MSC/MSE
Within Samples	SSE	$v_2=n-c$	$MSE=SSE/(n-c)$	
Total	SST	$n-1$		

SST - Total sum of squares of variations

SSC - Sum of squares between samples (columns)

SSE - Sum of squares within samples (rows)

MSC - Mean sum of squares between samples

MSE - Mean sum of squares within samples

Note: The same procedure for analysis of variance is applicable for both the equal and unequal sample sizes.

Example 1: To assess the significance of possible variation in performance in a certain test between the grammar schools of a city, a common test was given to a number of students taken at random from the senior fifth class of each of the four schools concerned. The results are given below. Make an analysis of variance of data.

Schools			
A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution:

Sample I		Sample II		Sample III		Sample IV.	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	64	14	196	6	36	16	256
7	49	4	16	8	64	15	225
ΣX_1 =45	ΣX_1^2 =421	ΣX_2 =50	ΣX_2^2 =58	ΣX_3 =60	ΣX_3^2 =824	ΣX_4 =65	ΣX_4^2 =875

The sum of all the items of various samples

$$= \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4$$

$$= 45 + 50 + 60 + 65 = 220.$$

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(220)^2}{20} = \frac{48400}{20} = 2420.$$

$$\text{Total sum of the squares} = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N}$$

$$= 421 + 558 + 824 + 875 - 2420$$

$$= 2678 - 2420 = 258.$$

Sum of squares between the samples:

$$= \frac{(\Sigma X_1)^2}{N} + \frac{(\Sigma X_2)^2}{N} + \frac{(\Sigma X_3)^2}{N} + \frac{(\Sigma X_4)^2}{N} - \frac{T^2}{N}$$

$$= \frac{(45)^2}{5} + \frac{(50)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420$$

$$= \frac{12350}{5} - 2420 = 2470 - 2420 = 50.$$

Mean sum of squares between the samples is $\frac{50}{4-1} = 16.7$.

$$v_1 = (\text{No. of samples} - 1 = 4 - 1 = 3)$$

Sum of squares within samples = Total sum of squares - Sum of squares between samples

$$= 258 - 50 = 208.$$

$$v_2 = \text{Total No. of items} - \text{No. of samples}$$

$$= 20 - 4 = 16.$$

Mean sum of squares within the samples

$$= \frac{208}{20-4} = \frac{208}{16} = 13.$$

ANOVA Table

Source of Variation	Sum of Squares	Degrees of freedom	Mean square
Between Samples	50	3	16.7
Within Samples	208	16	13
Total	258	19	

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{16.7}{13} = 1.285.$$

The table value of F for $v_1=3$ and $v_2=16$ at 5% level of significance = 3.24. The calculated value of F is less than the table value. Hence the difference is not significant. i.e., the samples could have come from the same universe.

Example 2: The three samples below have been obtained from normal populations with equal variances. Test the hypothesis that the sample means are equal.

8	7	12
10	5	9
7	10	13
14	9	12
11	9	14

The table value of F at 5% level of significance for $v_1=2$ and $v_2=12$ is 3.88.

Solution: Let us take the hypothesis that there is no significant difference in the means of three samples. Using analysis of variance technique.

	X_1	X_2	X_3
	8	7	12
	10	5	9
	7	10	13
	14	9	12
	11	9	14
Total	50	40	60
$\bar{X}$	10	8	12

$$\bar{X} = \frac{10 + 8 + 12}{3} = 10.$$

Variance between samples

$(\bar{X}_1 - \bar{X})^2$	$(\bar{X}_2 - \bar{X})^2$	$(\bar{X}_3 - \bar{X})^2$
0	4	4
0	4	4
0	4	4
0	4	4
0	4	4
0	20	20

Sum of squares between samples = $0 + 20 + 20 = 40$.

Mean squares between samples = $\frac{40}{3-1} = \frac{40}{2} = 20$.

Variance within samples

X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$	X_3	$(X_3 - \bar{X}_3)^2$
8	4	7	1	12	0
10	0	5	9	9	9
7	9	10	4	13	1
14	16	9	1	12	0
11	1	9	1	14	4
	30		16		14

Sum of squares within samples = $30 + 16 + 14 = 60$.

Mean squares within samples = $\frac{60}{15-3} = \frac{60}{12} = 5$.

ANOVA Table

Source of variation	Sum of squares	v	Mean square
Between	40	2	20
Within	60	12	5
Total	100	14	

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{20}{5} = 4$$

For $v_1 = 2$ and $v_2 = 12$, $F_{0.05} = 3.88$

The calculated value of F is greater than the table value. Hence the hypothesis is rejected. So there is significant difference in the sample means.

Analysis of variance in two - way classification model

In a two way classification the data are classified according to two different criteria or factors. The analysis of variance table takes the following form:

Source of variation	Sum of Squares	Degrees of freedom	Mean Sum of squares	Ratio of F
Between samples	SSC	$(c - 1)$	$MSC = SSC / (c - 1)$	MSC / MSE
Between Rows	SSR	$(r - 1)$	$MSR = SSR / (r - 1)$	MSR / MSE
Residual error	SSE	$(c - 1)(r - 1)$	$MSE = SSE / ((r - 1)(c - 1))$	
Total	SST	$n - 1$		

SSC = Sum of squares between columns

SSR = Sum of squares between rows

SSE = Sum of squares due to error

SST = Total sum of squares

The sum of squares for the source 'Residual' is obtained by $SSE = SST - [SSC + SSR]$.

The total number of degrees of freedom = $cr - 1$

Where C refers to number of columns and r refers to number of rows

Number of degrees of freedom between columns = $C - 1$.

Number of degrees of freedom between rows = $r - 1$.

Number of degrees of freedom for residual = $(C - 1)(r - 1)$

The F values of are calculated as follows:

$$F_{(v_1, v_2)} = \frac{MSC}{MSE} \text{ where } v_1 = c - 1 \text{ and}$$

$$v_2 = (c - 1)(r - 1)$$

$$F_{(v_1, v_2)} = \frac{MSR}{MSE} \text{ where } v_1 = r - 1 \text{ and}$$

$$v_2 = (c - 1)(r - 1)$$

It should be carefully noted that v_1 may not be same in both cases - in one case $v_1 = c - 1$ and another case $v_1 = r - 1$.

The calculated values of F are compared with the table values. If calculated value of F is greater than the table value at pre-assigned level of significance, the null hypothesis is rejected, otherwise accepted.

In two way classification, 'Residual' is the measuring rod for testing significance. It represents the magnitude of variation due to forces called 'chance'.

Example 1: A tea company appoints four salesmen A, B, C and D and observes their sales in three seasons - summer, winter and monsoon. The figures (in lakhs) are given in the following table:

Seasons	Salesmen				Season's Total
	A	B	C	D	
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Salesmen's Totals	90	93	81	96	360

Carry out an analysis of variance

Solution: The above data are classified according to criteria (i) salesmen and (ii) seasons. In order to simplify calculations we code the data by subtracting 30 from each figure. The data in the coded form are:

Seasons	Salesmen				Season's Total
	A	B	C	D	
Summer	6	6	-9	5	8
Winter	-2	-1	1	2	0
Monsoon	-4	-2	-1	-1	-8
	0	3	-9	6	Grand Total T=0

$$\text{Correction factor} = \frac{T^2}{N} = \frac{0}{12} = 0.$$

Sum of squares between salesmen

$$= \frac{0^2}{3} + \frac{(3)^2}{3} + \frac{(-9)^2}{3} + \frac{6^2}{3} - \frac{T^2}{N}$$

$$= 0 + 3 + 27 + 12 + 0 = 42.$$

$$v = 4 - 1 = 3.$$

Sum of squares between seasons

$$= \frac{8^2}{4} + \frac{0^2}{4} + \frac{(-8)^2}{4} - \frac{T^2}{N}$$

$$= 16 + 0 + 16 - 0 = 32$$

$$v = 3 - 1 = 2$$

Total sum of squares

$$= (6)^2 + (-2)^2 + (-4)^2 + (6)^2 + (-1)^2 + (-2)^2 + (-9)^2 + (1)^2 + (-1)^2 + (5)^2 + 2^2 + (-1)^2 - \frac{T^2}{N}$$

$$= 36 + 4 + 16 + 36 + 1 + 4 + 81 + 1 + 1 + 25 + 4 + 1 - 0$$

$$= 210 - 0 = 210.$$

$$v = (12 - 1) = 11.$$

ANOVA Table

Source of variation	Sum of squares	d.f.	Mean square
Between columns (Salesmen)	42	3	14
Between rows (Seasons)	32	2	16
Residual	136	6	22.67
Total	210	11	

Let us take the hypothesis that there is no difference between the sales of salesmen and of seasons.

First compare the salesmen variance estimate with the residual variance estimate; thus

$$F = \frac{22.67}{14} = 1.619$$

The table value of F for $v_1=3$ and $v_2=6$ at 5% level of significance is 4.76.

The calculated value is less than the table value and we conclude that the sales of different salesmen do not differ significantly.

Let us compare the season variance estimate with the residual variance estimate; thus

$$F = \frac{22.67}{16} = 1.417$$

The critical value of F for $v_1 = 2$ and $v_2 = 6$ at 5% level of significance is 4.76. The calculated value is less than this and hence there is no significant difference in the seasons as far as the sales are concerned.

Thus the test shows that the salesmen and the seasons are alike so far as the sales are concerned.

Example 2: The following table gives the number of units of production per day turned out by four different types of machines;

Employee	Type of machines			
	M ₁	M ₂	M ₃	M ₄
E ₁	40	36	45	30
E ₂	38	42	50	41
E ₃	36	30	48	35
E ₄	46	47	52	44

Using Analysis of variance (i) test the hypothesis that the mean production is the same for the four machines and (ii) test the hypothesis that the employees do not differ with respect to mean productivity.

Solution: Let us take the hypothesis (i) that the mean production does not differ for the four machines and (ii) the employees do not differ with respect to mean productivity.

Subtracting 40 from the given values to simplify calculation, we get:

Employee	Type of machines				
	M ₁	M ₂	M ₃	M ₄	Total
E ₁	0	-4	5	-10	-9
E ₂	-2	2	10	1	11
E ₃	-4	-10	8	-5	-11
E ₄	6	7	12	4	29
Total	0	-5	35	-10	20

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(20)^2}{16} = 25.$$

Sum of squares between machines:

$$= \frac{0^2}{4} + \frac{(-5)^2}{4} + \frac{(-35)^2}{4} + \frac{(-10)^2}{4} - \frac{T^2}{N}$$

$$= 0 + 6.25 + 306.25 + 25 - 25 = 312.5$$

$$v = 4 - 1 = 3.$$

Sum of squares between employees:

$$= \frac{(-9)^2}{4} + \frac{(11)^2}{4} + \frac{(-11)^2}{4} + \frac{(29)^2}{4} - \frac{T^2}{N}$$

$$= 20.25 + 30.25 + 30.25 + 210.25 - 25 = 266.$$

$$v = 4 - 1 = 3$$

$$\text{Total sum of squares} = (0)^2 + (-2)^2 + (-4)^2 + (6)^2 + (-4)^2 + 2^2 + (-10)^2 + 7^2 + 5^2 + (10)^2$$

$$+ 8^2 + (12)^2 + (-10)^2 + 1^2 + (-5)^2 + 4^2 - \frac{T^2}{N}$$

$$= 0 + 4 + 16 + 36 + 16 + 4 + 100 + 49 + 25 + 100$$

$$+ 64 + 144 + 100 + 1 + 25 + 16 - 25 = 675$$

$$\text{Residual} = 675 - (312.5 + 266) = 96.5.$$

ANOVA Table

	Sum of squares	v	Mean square	F
Between columns	312.5	3	101.17	$\frac{104.17}{10.72} = 9.72$
Between Rows	266.0	3	88.67	$\frac{88.67}{10.72} = 8.27$
Residual	96.5	9	10.72	
Total	675	15		

For $v=(3,9)$, $F_{0.05} = 3.86$. The calculated value of F are more than the table value at 5% level of significance. So the hypothesis is rejected. Hence the mean production is not the same for the four machines. Also employees differ with respect to mean productivity.

Exercise

- The following table gives the yields on 15 sample plots under three varieties of seed:

	A	B	C
20	18	25	
21	20	28	
23	17	22	
16	15	28	
20	25	32	

We have to find out if the average yields of land under different variances show significant differences.

- The following figures relate to the number of units sold in five different areas by four salesmen:

Area	Number of units			
	A	B	C	D
1	80	100	95	70
2	82	110	90	75
3	88	105	100	82
4	85	115	105	88
5	75	90	80	65

Is there a significant difference in the efficiency of these salesmen?

Table value of $F_{0.05}$ for $v_1 = 3$, $v_2 = 12$ is 3.24.

3. To study the performance of three detergents and three different water temperatures, the following 'Whiteness' readings were obtained with specially designed equipment:

Water Temp.	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
Hot water	54	46	58

Perform a two-way analysis of variance, using 5% level of significance (Given $F_{0.05}$ for $v_1=2$, $v_2=3$ is 6.94)

4. To test the significance of the variation of the retail prices of a commodity in three principal cities, Bombay, Calcutta and Delhi, four shops were chosen at random in each city and prices observed in rupees were as follows:

Bombay :	16	8	12	14
Calcutta :	14	10	10	6
Delhi :	4	10	8	8

Do the data indicate that the prices in the three cities are significantly different?

5. The following table gives monthly sales (in thousand rupees) of a certain firm in three states by its four salesmen:

States	Salesmen			
	I	II	III	IV
A	6	5	3	8
B	8	9	6	5
C	10	7	8	7

Set up the analysis of variance table and test whether there is any significant difference (i) between sales by the firm salesmen and (ii) between sales in the three states.

6. In a certain factory, production can be accomplished by four different workers on five different types of machines. A simple study, in context of a two way design without repeated values is being made with two fold objectives of examining whether the four workers differ with respect to mean productivity and whether the mean productivity is the same for the five different machines. The researcher involved in their study reports while analysing the gathered data as under:

- (i) Sum of squares for variance between machines = 35.2
- (ii) Sum of squares for variance between workmen = 53.8
- (iii) Sum of squares for total variance = 174.2

Set up ANOVA table for the given information and draw the inference about variance at 5% level of significance.

10.5 - Three criteria of classification - Latin Square

Latin squares are very extensively used in agricultural trials in order to eliminate fertility trends in two directions simultaneously. The data are classified according to the different criteria and are arranged in a square known as Latin Square. The term Latin Square takes its name from a figure of mathematical puzzle. It is an arrangement of the letters ie., varieties in a square in such a way that each letter (variety) occurs once and only once in each row and each column. For example, in a 5 by 5 (5x5) Latin Square, where data may be classified according to rows, columns and varieties, varieties being represented by letters A,B, C,D,E, the arrangement may be as below:

	1	2	3	4	5
1	A	B	C	D	E
2	B	E	A	C	D
3	C	D	B	E	A
4	D	C	E	A	B
5	E	A	D	B	C

The total number of possibilities are 161, 280.

Significance of Latin Square

Latin Square designs have a wide variety of applications in experimental work. They are used in agriculture, industrial, laboratory field, educational, medical, marketing and sociological experimentation.

A Latin Square design may well be useful when we wish to remove from an analysis of data the effect of a factor which we are not interested in but which is known to be significant.

Assumption in the Analysis of Latin Square

The Latin Square model assumes that interactions between treatments and row and column groupings are non-existent. Since each treatment occurs only once in each row or columns, if interactions are present, it is possible for them to cause an apparently significant difference between treatments. This is one of the reasons why it is important to choose rows and columns of a particular Latin Square in a random way.

Steps in Constructing Latin Square

1. Compute the correction factor by squaring the grand total and dividing it by the number of observations.

2. Compute the total sum of squares by adding the squares of the individual observations and subtracting the correction factor.
3. Compute the row sum of squares by adding the squares of the row sums, dividing by the number of items in a row, and subtracting the correction factor.
4. Compute the column sum of squares by adding the squares of the column sums, dividing by the number of items in a column, and subtracting the correction factor.
5. Compute the 'treatment' sum of squares by summing the squares of the treatment sums dividing by the number of treatments and subtracting the correction factor.
6. Compute the remainder sum of squares by subtracting the sum of 3, 4 and 5 from 2.
7. Enter these sums of squares in an analysis of variance table and compute the various mean squares.

Analysis of Variance Table

Source of variance	S.S.	Degrees of freedom	Mean Square
Rows	SSR	$n-1$	$MSR=SSR/(n-1)$
Columns	SSC	$n-1$	$MSC=SSC/(n-1)$
Treatments	SST	$n-1$	$MST=SST/(n-1)$
Residual or error	SSE	$(n-1)(n-2)$	$MSE=SSE/(n-1)(n-2)$
Total	TSS	n^2-1	

8. The last step is to calculate F by comparing the treatment mean square with the remainder mean square.

Example 1: Five varieties of wheat A, B, C, D and E were tried. The gross size of the plot was 18 feet x 12 feet the net plot being 14 feet x 18 feet. Thus the whole experiment occupied an area 90 feet x 110 feet. The plan, the varieties sown in each plot and yields and obtained in kg are given in the following table:

B 90	E 80	C 134	A 112	D 92
E 85	D 84	B 70	C 141	A 82
C 110	A 90	D 87	B 84	E 69

A	C	E	D	B
81	125	85	76	72
D	B	A	E	C
82	60	94	85	88

Carry out an analysis of variance. What inference can you draw from the data given?

Solution: It is a 5 x 5 Latin Square. Let us subtract 90 from each value, the table value will be as follows:

Rows	Columns					Total
	1	2	3	4	5	
1	0	-10	44	22	2	58
2	-5	-6	-20	51	-8	12
3	21	0	-3	-6	-21	-9
4	-9	35	-5	-14	-18	-11
5	-8	-30	4	-5	-2	-41
Total	-1	-11	20	48	-47	9

Squares					
	1	2	3	4	5
1	0	100	1936	484	4
2	25	36	400	2601	64
3	441	0	9	36	441
4	81	1225	25	196	324
5	64	900	16	25	4
	611	2261	2386	3342	837

$$\Sigma X^2 = 611 + 2261 + 2386 + 3342 + 837 = 9437$$

Treatment	1	2	3	4	5	Total
A	-9	0	4	22	-8	9
B	0	-30	-20	-6	-18	-74
C	21	35	44	51	-2	149
D	-8	-6	-3	-14	2	-29
E	-5	-10	-5	-5	-21	-46

$$N = 5 \times 5 = 25.$$

$$\begin{aligned} \text{Total sum of squares} &= \sum X^2 - \frac{T^2}{N} \\ &= 9437 - \frac{(9)^2}{25} \end{aligned}$$

$$= 9437 - 3.24 = 9433.8$$

$$\text{Total degrees of freedom} = 25 - 1 = 24.$$

2) Sum of squares between columns

$$\begin{aligned} &= \frac{1}{5} [(-1)^2 + (-11)^2 + (20)^2 + (48)^2 + (-47)^2] - \frac{T^2}{N} \\ &= \frac{1}{5} [1 + 121 + 400 + 2304 + 2209] - 3.24 \\ &= 1007 - 3.24 = 1003.76 \text{ or } 1003.8. \end{aligned}$$

$$v = r - 1 = 4$$

3) Sum of squares between rows

$$\begin{aligned} &= \frac{1}{5} [(58)^2 + (12)^2 + (-9)^2 + (-11)^2 + (-41)^2] - \frac{T^2}{N} \\ &= \frac{1}{5} (3364 + 144 + 81 + 121 + 1681) - 3.24 \\ &= 1078.2 - 3.24 = 1074.96 \text{ or } 1075. \end{aligned}$$

$$v = r - 1 = 4.$$

4) Sum of squares between treatments

$$\begin{aligned} &= \frac{1}{5} [9^2 + (-74)^2 + (149)^2 + (-29)^2 + (-46)^2] - \frac{T^2}{N} \\ &= \frac{1}{5} [81 + 5476 + 22201 + 841 + 2116] - 3.24 \\ &= \frac{1}{5} [30715] - 3.25 \end{aligned}$$

$$= 6143 - 3.24 = 6139.76 \text{ or } 6139.8$$

5) Degrees of freedom between treatments = $t - 1 = 5 - 1 = 4$.

6) Residual = $9433.8 - [1003.8 + 1075 + 6139.8]$

$$= 1215.2$$

$$v = 4 \times 3 = 12.$$

ANOVA Table

Source	v	S.S.	M.S.	F - ratio observed	F _{0.05}
Rows	4	1075	268.75	$2.65 = \left(\frac{268.75}{101.27} \right)$	3.26
Columns	4	1003.8	250.95	$2.48 = \left(\frac{250.95}{101.27} \right)$	3.26
Varieties	4	6139.8	1534.95	$15.15 = \left(\frac{1534.95}{101.27} \right)$	3.26
Error	12	1215.2	101.27		
Total	24	9433.8			

The F ratios for rows and columns are not significant at 5% level while that for varieties is very highly significant. There are no significant differences between rows and columns shows that the Latin square arrangement has not been advantageous.

Example 2: Analyse the variance in the following Latin square.

A	C	B
8	18	9
C	B	A
9	18	16
B	A	C
11	10	20

Solution: Subtracting each by 9 we have the table:

	I	II	III	IV
1	-1	9	0	8
2	0	9	7	16
3	2	1	11	14
T	1	19	18	38

T=38.

Squares

	Total		
1	81	0	82
0	81	49	130
4	1	121	126
Total	5	163	170

$$N = 3 \times 3 = 9.$$

$$\begin{aligned} \text{Total sum of squares} &= \sum x^2 - \frac{T^2}{N} \\ &= 338 - \frac{(38)^2}{9} \\ &= 338 - 160.4 = 177.6. \end{aligned}$$

$$\text{Total degrees of freedom} = 9 - 1 = 8.$$

Sum of squares between columns

$$\begin{aligned} &= \frac{1}{3} [1^2 + 19^2 + 18^2] - \frac{T^2}{N} \\ &= 228.7 - 160.4 = 68.3. \end{aligned}$$

$$v = c - 1 = 3 - 1 = 2.$$

Sum of squares between rows

$$\begin{aligned} &= \frac{1}{3} (8^2 + 16^2 + 14^2) - \frac{T^2}{N} \\ &= 172 - 160.4 = 11.6 \end{aligned}$$

$$v = r - 1 = 3 - 1 = 2$$

Treatment	1	2	3	Total
A	-1	7	1	7
B	0	9	2	11
C	9	0	11	20

Sum of squares between treatments

$$= \frac{1}{3} [7^2 + 11^2 + 20^2] - \frac{T^2}{N}$$

$$= \frac{1}{3} (570) - 160.4 = 190 - 160.4 = 29.6.$$

$$v = t - 1 = 3 - 1 = 2.$$

$$\text{Residual} = 177.6 - (68.3 + 11.6 + 29.6) = 68.1$$

$$v = 1 \times 2 = 2$$

ANOVA Table

Source	v	S.S.	M.S.	F - ratio observed	F _{0.05}
Rows	2	11.6	5.8	$\frac{5.8}{34.1} = 0.17$	19
Columns	2	68.3	34.15	$1 = \left(\frac{34.15}{34.1} \right)$	19
Varieties	2	29.6	14.8	$0.43 = \left(\frac{14.8}{34.1} \right)$	19
Error	2	68.1	34.1		
Total	8	177.6			

All the calculated values of F are less than the table value. $\therefore$ The variations are not significant.

Exercise

1. Analyse the variance in the following Latin squares

(i)

B	C	D	A
20	17	25	34
A	D	C	B
23	21	15	24
D	A	B	C
24	26	21	19
C	B	A	D
26	23	27	22

(ii)

D 38	E 37	C 35	B 36	A 34
B 32	D 34	E 34	A 36	C 33
C 33	A 33	B 33	D 34	E 33
E 32	B 34	A 33	C 33	D 34
A 31	C 32	D 31	E 31	B 30

Source	v	S.S.	M.S.	F-ratio observed	F _{0.05}
Rows	2	11.6	5.8	$\frac{5.8}{34.1} = 0.17$	19
Columns	2	68.3	34.15	$1 = \left(\frac{34.15}{34.1} \right)$	19
Variates	2	29.6	14.8	$0.43 = \left(\frac{14.8}{34.1} \right)$	19
Error	2	68.1	34.1		
Total	8	177.6			

All the calculated values of F are less than the table value. The variations are not significant.

Analyse the variance in the following Latin squares

(i)

B 20	A 23	D 24	C 26
A 23	D 21	B 21	A 23
D 24	B 21	A 23	B 21
C 26	A 23	B 21	D 24

AREA TABLE FOR STANDARD NORMAL DISTRIBUTION

$$\text{Area Under the normal curve } \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2853
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4441
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4545
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4633
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4706
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4767
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4216
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4063	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4986	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

TABLE OF ORDINATES FOR THE STANDARD NORMAL CURVE

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034

TABLE FOR STUDENT'S T-DISTRIBUTION

$\frac{P}{d.f}$	.05	.02	.01
1	12.706	31.821	63.657
2	4.303	6.965	9.925
3	3.182	4.541	5.841
4	2.776	3.747	4.604
5	2.571	3.365	4.032
6	2.447	3.143	3.707
7	2.365	2.998	3.499
8	2.306	2.896	3.355
9	2.262	2.821	3.250
10	2.228	2.764	3.169
11	2.201	2.718	3.106
12	2.179	2.681	3.055
13	2.160	2.650	3.012
14	2.145	2.624	2.977
15	2.131	2.602	2.947
16	2.120	2.583	2.921
17	2.110	2.567	2.898
18	2.101	2.552	2.878
19	2.093	2.539	2.861
20	2.086	2.528	2.845
21	2.080	2.518	2.831
22	2.074	2.508	2.819
23	2.069	2.500	2.807
24	2.064	2.492	2.797
25	2.060	2.485	2.787
26	2.056	2.479	2.779
27	2.052	2.473	2.771
28	2.048	2.467	2.763
29	2.045	2.462	2.756
inf	1.960	2.326	2.576

TABLE FOR χ^2 - DISTRIBUTION

$\frac{P}{d.f}$	.05	.02	.01
1	3.841	5.214	6.635
2	5.991	7.824	9.210
3	7.851	9.837	11.341
4	9.488	11.668	13.277
5	11.07	13.388	15.086
6	12.592	15.033	16.812
7	14.067	16.622	18.475
8	15.507	18.168	20.090
9	16.919	19.679	21.666
10	18.307	21.161	23.209
11	19.675	22.618	24.725
12	21.026	24.054	26.217
13	22.362	25.472	27.688
14	23.685	26.873	29.141
15	24.996	28.295	30.578
16	26.296	29.633	32.000
17	27.587	30.995	33.409
18	28.869	32.346	34.805
19	30.144	33.687	36.191
20	31.410	35.020	37.566
21	32.671	36.343	38.932
22	33.924	37.659	40.289
23	35.172	38.968	41.638
24	36.415	40.270	42.980
25	37.652	41.566	44.314
26	38.885	41.856	45.642
27	40.113	44.140	46.963
28	41.337	45.419	48.278
29	42.557	46.693	49.588
inf	43.773	47.962	50.892

TABLE FOR F- DISTRIBUTION

F-Test 5% points

v2 \ v1	1	2	3	4	5	6	8	12	24
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0
2	18.51	19.00	19.61	19.25	19.30	19.33	19.37	19.41	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89