PROBABILITY & STATISTICS

(DMCA208)

(MCA)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

UNIT -I

PROBABILITY AND DISTRIBUTIONS

Syllabus:

Definition of probability , classical and relative frequency approach to probability, merits and demerits of these approaches, random experiment sample point and sample space, definition of an event, operation of events, properties of probability based on axiomatic approach, addition theorem for n events.

Conditional Probability, multiplication rule of probability for n events , independence of events, Bayes' theorem and its applications (with examples in real life)

Notion of a random variable, distribution function and its properties, discrete random variable, probability mass function, continuous random variable, probability density function, mathematical expectation, Binomial distribution , Poisson distribution , simulating a Discrete distribution, Exponential distribution, Normal distribution, We bull distribution and Reliability.

Objective:

This lesson is prepared in such a way that after studying the material the student is expected to have a thorough comprehension of the concept of "Probability" – the breath of any statistical investigation and analysis. The student would be equipped with theoretical as well as practical aspects of probability of an event or combination of events.

The student is also expected to have a clear notion of probabilities of independent events. Its application in making decision about conditional events and the principle in finding Bayes theorem in assessing the performance of devices with built in structure probability

The student will be having a clear comprehension of the theory and the practical utility about the concepts of random variable , distribution function its properties, discrete random variable, probability mass function , continuous random variable, Probability density function, expectation, Binomial distribution , Poisson distribution , simulating a Discrete distribution, Exponential distribution, Normal distribution, Weibull distribution and Reliability.

Structure of the lesson:

1.1 Introduction:

Frequently we come across certain statements that are not always true and not always false. For instance the forecasting of weather in news bulletins, announcements about arrivals and departures of trains in a railway station, the results of pre poll surveys in general elections etc. In all these examples we see an element of uncertainty associated with them that would prevent us from taking an appropriate decision. Therefore, if there is a method of expressing uncertainty in numerical quantity, depending on the method magnitude of the numerical quantity one can decide whether or not to go ahead with a decision.

The word probability or chance is used commonly in day to day life. For example the chances of India and Pakistan winning the world cup cricket, before the start of the game are equal i.e., 50:50. It is likely that Mr. Kishore may not come for taking his class today. We often say that it is very probable that it will rain tomorrow. Probably I will not come to tea party tomorrow. All these terms chance, likely, probable, etc. convey the same meaning i.e., that event is not certain to take place. In other words, there is uncertainty about the happening of the event. Hence, in each of these cases we talk about chance or probability which is taken to be a quantitative measure of certainty. In this lesson we discuss at length the notion of probability, the various advancements in its definition, some standard results along with specific applications some standard results along with specific applications.

1.2 Basic Prerequisites :

In this section we present some concepts to explain probability.

1.2.1 Definition :

Random experiment.An experiment whose result is not known with certainty unless the experiment is performed completely.

Example 1. Throwing of coin or die, getting head or 1 or 2 or ….or

6.

2. Drawing of cards from a well shuffled pack of cards.

3. An agricultural experiment to determine the effects of

Fertilizers on yield of a commodity.

4. Winning or losing a match.

In all these examples some action is performed with an intended result. But the expected result may or may not happen in fact we experience many random experiments in our daily observations.

1.2.2 **Definition:** Sample space: In a random experiment we cannot say exactly guess outcome of the action, with some enlightened vision we can say the various possible results for the experiment, the set of all possible outcomes of a random experiment without any omission is called sample space. In the examples of the definition 1.2.1 the following sets are sample spaces respectively.

{Head, Tail}, {1,2,3,4,5,6}

{52 cards}

{Agricultural land}

{Win, lose, drawn}

Sample space is similar to the universal set in set theory and is denoted by 2. The elements of Ω one called sample point which one also called simple events. Combination of simple events is called an event. That is sub sets of Ω are called events. For example in a dice throwing example the single ton sets { 1},{ 2 },{ 3 }, {4 },{ 5 }, { 6 } are called sample points. The subset {1,3,5} is an event.

Denoting and getting an odd number and the set {3, 6} denoted the event of getting a multiple of 3.

Hence we can think of a parallel between set theory and events in sample space. If A and B are any two subsets of Ω

$A^1 = \Omega$ - A is called complimentary event to A.

AUB = Occurrence of either of the events.

A∩B = Occurrence of both the events A,B.

Also if A∩B = Ø the null set then A,B are specifically called mutually exclusive events.

1.3 Relative frequency approach:

In a trial is repeated a number of times under essentially homogeneous and identical conditions, then the limiting value 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 the probability of happening of the event. That is, if n trials an event E happens m times, then the probability 'P' of the happening of E is given by

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

The relative frequency approach to the definition of probability is basically a limit of a sequence. Hence, unless the sequence is convergent, we cannot get the probability. Even if it is convergent one may have to do a number of repetitions of the experiment, which may be a costly affair, therefore this remains more of a theoretical proportion than a practically adaptable definition.

1.4 Classical definition of probability:

In a random experiment if a trial results in n exhaustive, mutually exclusive and equally likely cases and m of them are favorable to the happening of an event E, than the probability 'p' of happening of event E is given by the ratio $\frac{m}{n}$.

Example 1.4.1: The sample in a die throwing experiment is Ω = { 1,2,3,4,5,6 }.Suppose if we are interested in getting a prime number . Then the set {2, 3, 5} is the interested event say E. Here Ω contains 6 elements and E contains three elements i.e., m=3, n=6. According to classical definition the probability of the event E is p (E) =3/6.

In this definition, if the number of points in the sample space is not finite, if the elements of Ω are not equally likely we cannot use classical definition. It hints that the elements of Ω shouldhave equal chance of happening which in turn means that they should have the same probability of occurrence. That is the notion of probability in classical approach. Hence this approach is not totally admissible. Overcoming all the demerits of relative frequency approach and classical approach, probability is defined in axiomatic approach by A.N. Kolmogorov in the early part of 20thcentury. We explain this approach in section 4.5

1.5 Axiomatic definition of probability:

In axiomatic approach to probability theory, the probability is defined as a function, which is defined on events. In other words it is a rule which associates certain real number P (A) to each event A and satisfies the following three axioms.

Axiom 1: p (A) ≥ 0, i.e., the probability of every event is non-negative.

Axiom 2: P (Ω) =1, i.e., the probability of a certain event is 1.

Axiom 3: If $A_1, A_2, \ldots\ldots A_n$ are finite number of disjoint events of Ω then

P ($A_1UA_2UA_3U$----) =P (A_1) +P (A_2) +P (A_3) + ----.

i.e.,$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)\,, where\ A_i \cap A_j = \emptyset \forall i \neq j$

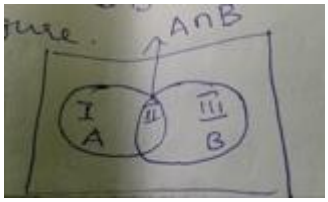
i.e., the probability of union of disjoint events is the sum of probabilities of the events themselves.

1.6 Addition Theorem:

For any two events A and B

P (AUB) =P (A) +P (B)-P (A∩B).

Proof: Let the events A and B be represented as the sets shown in the figure.



The regions I, II, III in the figure are mutually disjoint.

Also I U II U III = AUB.

I U II = A.

II U III = B.

By the third rule in the axiomatic definition of probability we get the following identities.

P (AUB) = P (I) +P (II) +P (III) ----- (1)

P (A) = P (I) +P (II) ------ (2)

P (B) = P (II) +P (III) ------ (3)

Subtracting the sum of equations (2) and (3) from (1) we get

P (AUB) –P (A) –P (B) = -P (II) ----- (4)

But region II is A∩B.

Equation (4) becomes

P (AUB) = P (A) +P (B) –P (A∩B).

1.7 Examples :

1. What is the Probability of getting a total of 7 in a single throw with two dice?

Sol: Since the die contains six sides when two dice are thrown the sample space contains 36 points given by

$\Omega = \{S_1 \times S_2\} = \{S_1=\{1, 2, 3, 4, 5, 6\} \times S_2 = \{1, 2, 3, 4, 5, 6\}\} = 6X6=36.$

The exhaustive number of events is n=6.

Also when two dice are thrown a total of 7 is obtained with the following pairs when we want $S_1+S_2 = 7$, i.e., (1, 6), (2, 5), (4, 3), (5, 2), (6, 1).

The events which are favorable of getting a total of 7 when two dice are thrown is m=6. Hence by classical definition of probability $= \frac{6}{36} = \frac{1}{6}$.

2. What is the probability of throwing more than 7 is equal to that of throwing less than 7. When two dice are thrown in a single trial?

Sol : In a dice throwing experiment when two dies one thrown the sample space contains 36 pairs of points hence the total or exhaustive number of events are n=36.

Out of (36) possible pairs the following pairs give the totals written against the respective group of pairs.

(1, 1): 2

(1, 2); (2, 1): 3

(1,3), (2,2),(3,1): 4

(1,4),(2,3),(3,2),(4,1): 5

(1,5),(2,4),(3,3),(4,2),(5, 1): 6

(2,6),(3,5),(4,4),(5,3),(6,2): 8

(3,6),(4,5),(5,4),(6, 3): 9

(4,6),(5,5),(6,4): 10

(5,6),(6,5): 11

(6,6):12

In the above the number of pairs that gives a total of 2 or 3 or 4 or 5 or r that is, a total of less than 7 is 15. Hence probability of getting a total less than 7 is 15/36. Similarly the total number of pairs to get a total of 8 or 9 or 10 or 11 or 12. That is, a total of more than 7 is 15.

The probability of getting a total of more than 7 is also 15/36.

3. In three coins are tossed. Find the probability of getting

(i) Three heads

(ii) Two heads

(iii) No heads

Sol: (i) getting three heads when three coins one tossed denotes the event E than.

E= {H H H} i.e., m=1

Sample space contains s= {HHH,HHT,THH,HTH,HTT,THT,TTH,TTT}

P(E) = m/n=1/8.

(ii) The events which are favorable of getting two heads is given by

E = {HHT,THH,HTH} i.e. m=3

And the sample space defined as S= {HHH,HHT,THH,HTH,HTT,THT,TTH,TTT } i.e., n=8.

P(E) = m/n=3/8.

(iii) The probability of getting no heads, than the event of getting no head when there coins are tossed is given by

E = {TTT} i.e., m=1.

And n = 8.

P(E) = m/n = 1/8.

4. Find the probability getting 2 diamonds. If we draw 2 cards at random from a packet of 52 cards.

Solution: In a pack of cards there will be 52 cards and 2 cards can be chosen m 52_{c_2} ways.

No.of exhaustive cases = 52_{c_2} = 1326=n

There are 13_{c_2} ways to choose 2 diamonds. Since there are 13 diamonds.

Number of favorable cases 13_{c_2} = 78=m.

Hence the required probability =m/n=78/1326=1/17.

5. Three cards are drawn from a pack of 52 cards. Find the probability that (i) 3 one spades (ii) 2 spades and one diamond (iii)1 spade, 1diamond , 1 heart.

Solution: There are 52_{c_3} ways to draw 3 cards from 52 cards

Exhaustive number of cases n= 52_{c_3} =22100

Favorable number of cases m = 13_{c_3} = 286

Required probability = $\frac{286}{22100} = \frac{143}{11050} = \frac{11}{850}$

(ii) The number of favorable cases for getting 2 spades and one diamond is m = $13_{c_2} \times 13_{c_1} = 78$

The exhaustive number of cases n=22100.

Hence the required probability = $\frac{m}{n} = \frac{78}{22100} = \frac{39}{850}$

(iii) The number of favorable cases for setting 1 spade 1diamond and 1heart is m = $13_{C_1} \times 13_{C_1} \times 13_{C_1} = 2197$

The number of exhaustive cases n = 22100

The required probability = $\frac{m}{n} = \frac{2197}{22100} = \frac{169}{1700}$

6. Three light bulbs are chosen at random from 12 bulbs of which 5 are defective. Find the probability that

(i) All are defective

(ii) One is defective

(iii) Two are defective

Sol: since three light bulbs one chosen at random from 12 bulbs, the exhaustive number of cases are given by

$n = 12_{C_3} = 220$

(i) As there are 5 defectives and 3 bulbs are chosen at random in $5_{C_3} = 10$ Required probability = $\frac{m}{n} = \frac{10}{220} = \frac{1}{22}$

(ii) Since 3 bulbs are chosen at random out of 12 bulbs of which 5 are defective then the number of favorable cases are

$5_{C_1} \times 7_{C_2}$ = 105.

Required probability = $\frac{105}{220} = \frac{21}{44}$

(iii) Since 2 are defective out of 3 light bulbs, the number of favorable cases is $m = 5_{C_2} \times 7_{C_1} = 70$.

Required probability = $\frac{m}{n} = \frac{70}{220} = \frac{7}{22}$

7. A bag contains 5 red balls, 8 blue balls and 11 white balls three balls one drawn together from the box. Find the probability that

(i) One is red, one is blue and one is white

(ii) Two whites and one red, (iii) three white.

Sol: Number of exhaustive cases $n = 24_{C_3} = 2024$

(i) Number of favorable cases $5_{C_1}.8_{C_1}.11_{C_1} = 440 = m$

Required probability = $\frac{m}{n} = \frac{440}{2024} = \frac{55}{253}$

(ii) Number of favorable cases = $11_{C_2}.5c_1 = 275 = m$

Required probability = $\frac{m}{n} = \frac{275}{2024} = \frac{25}{184}$

(iii) Number of favorable cases $m = 11_{C_3} = 165$

Required probability = $\frac{m}{n} = \frac{165}{2024} = \frac{15}{184}$

8. What is the probability of drawing an ace from a well shuffled pack of 52 playing cards?

Sol: The number of exhaustive cases $n = 52_{C_1} = 52$

The number of favorable cases $m = 4_{C_1} = 4.$

Required probability = $\frac{m}{n} = \frac{4}{52} = \frac{1}{13}$

8.1 State whether the following probabilities are permissible

(i) $P(A_1) = \frac{-1}{2}, p(A_2) = \frac{1}{4}, p(A_3) = \frac{1}{4}, p(A_4) = 0$
$where, S = \{A_1, A_2, A_3, A_4\}$

$whereS = \{A_1, A_2, A_3\}$

(ii) $P(A_1) = \frac{1}{3}, P(A_2) = \frac{1}{3}, P(A_3) = \frac{1}{6},$

(iii) $S = \{A_1, A_2, A_3\}, P(A_1) = 0, p(A_2) = \frac{1}{2}, p(A_3) = \frac{1}{2}$

(iv) $S = \{A_1, A_2, A_3, A_4, A_5\}, p(A_1) = \frac{1}{5}, p(A_2) = \frac{1}{10}, p(A_3) = \frac{3}{10}, p(A_4) = \frac{1}{5}, P(A_5) = \frac{1}{5}$

Solution: (i) Cannot be permissible since p(A_1) is negative.

(ii) Not permissible since $P(A_1) + p(A_2) + P(A_3) = P(S) \neq 1.$

(iii) permissible since $P(A_1) + p(A_2) + P(A_3) = 1.$&

$P(A_1), p(A_2), P(A_3) \geq 0.$

(iv) permissible since $P(A_1) + p(A_2) + P(A_3) + p(A_4) + P(A_5) = 1.$

i.e $P(A_i) > 0 \, for i = 1,2,3,4,5.$

9. Among 150 students 80 are studying maths, 40 are studying physics and 30 are studying maths and physics if a student is chosen at random. Find the probability that the student.

(i) Studying maths or physics

(ii) Student studying neither maths nor physics

Sol : Let event A be student studying maths.

B be student studying physics

$$\therefore P(A)=\frac{80}{150}, P(B)=\frac{40}{150}, p(A\cap B)=\frac{30}{150}$$

(i) Probability that a student is studying maths or physics is

$$p(A\cup B)=p(A)+P(B)-P(A\cap B)$$
$$=\frac{80}{150}+\frac{40}{150}-\frac{30}{150}=\frac{9}{15}=\frac{3}{5}$$

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

10. Suppose A and B are any two events and P(A)=p1, P(B)=P2

$p(A\cap B)=P_3$ prove the following identities.

(i) P($P(\bar{A}\cup B)=\bar{1}-P_3$

(ii) $p(\bar{A}\cap\bar{B})=1-p_1-p_2+p_3$

(iii) $p\left(A\cap\bar{B}\right)=P_1-P_3$

(iv) $P(\bar{A}\cap B)=P_2-P_3$

(v) $P\left(\overline{A\cap B}\right)=1-P_3$

(vi) $P(\bar{A}\cup B)=(1-P_1+P_3)$

(vii) $P(\overline{A\cup\bar{B}})=1-P_1-P_2-P_3$

(viii) $P(\bar{A}\cap(A\cup B))=P_2-P_3$

(ix) $P(A\cup(\bar{A}\cap B))=P_1+P_2-P_3$

From the results of set we know that

(i)
$$\bar{A}\cup\bar{B}=\overline{A\cap B}$$
$$\therefore P(\bar{A}\cup\bar{B})=P(\overline{A\cap B})=1-P_3$$

(ii)
$$\bar{A}\cap\bar{B}=\overline{A\cup B}\Rightarrow P(\bar{A}\cap\bar{B})=P(\overline{A\cup B})$$
$$i.e., P(\bar{A}\cap\bar{B})=1-P_1-P_2+P_3(\because 1-P(\overline{A\cup B})=1-[P(A)+P(B)-P(A\cap B)]$$

(iii) $A\cap\bar{B}$ can be interpreted as follows

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

Also $A \cap \overline{B}$ and $A \cap B$, are mutually exclusive events

$$P(A) = P(A \cap \overline{B}) + P(A \cap B)$$
$$i.e,.P_1 = P(A \cap \overline{B}) + P_3 \Rightarrow P(A \cap \overline{B}) = P_1 - P_3$$

(iv) Consider
$$(\overline{A} \cap B) \cup (A \cap B) = B$$
$$ALSO, \overline{A} \cap B, A \cap B, ARE, MUTUALLYEXCLUSIVE$$
$$\therefore P((\overline{A} \cap B) \cup (A \cap B)) = P(B)$$
$$P(\overline{A} \cap B) + P(A \cap B) = P(B)$$
$$P(\overline{A} \cap B) + P_3 = P_2$$
$$\therefore P(\overline{A} \cap B) = P_2 - P_3.$$

(v) $P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - P_3$

(vi)
$$(\overline{A} \cup B) = A \cap \overline{B}$$
$$P(\overline{A} \cup B) = 1 - P(\overline{\overline{A} \cup B}) = 1 - P(A \cap \overline{B})$$
$$= 1 - (P_1 - P_3) FROM (iii)$$

(vii)
$$P(\overline{A \cup B}) = 1 - P(A \cup B)$$
$$= 1 - [P(A) + P(B) + -P(A \cap B)]$$
$$= 1 - [P_1 + P_2 + P_3] = 1 - P_1 - P_2 + P_3$$

(viii)
$$\overline{A} \cap (A \cup B)$$
$$= (\overline{A} \cap A) \cup (\overline{A} \cap B)$$
$$= \phi \cup (\overline{A} \cap B) = \overline{A} \cap B.$$

FROM IV we know that
$$P(\overline{A} \cap B) = P_2 - P_3$$
$$\therefore P(\overline{A} \cap (A \cup B)) = P_2 - P_3$$

(ix) Consider $A \cup (A \cap B)$

Appling addition law of probability to the sets A and $\overline{A} \cap B$, we get,

$$P(A \cup (\overline{A} \cap B)) = P(A) + P(\overline{A} \cap B) - P(A \cap \overline{A} \cap B)$$
$$= P(A) + P(\overline{A} \cap B)(SINCE, A, \overline{A} \cap B, AREDISJOINT$$
$$= P_1 + (P_2 - P_3) = P_1 + P_2 - P_3 (\because P(\overline{A} \cap B) = P_2 - P_3)$$

11. Two bolts are drawn form a box containing 4 good and 6 bad bolts find the probability that the second bolt is good if the first one is found to be bad ?

Ans : There are 4 good bolts and 6bad bolts

The probability that first one is bad is 6/10.

The probability that the second is good if first one is bad is

$$\frac{6}{10}.\frac{4}{9}=\frac{4}{15}$$

12. A class has 10 boys and 5 girls .three students are selected at random one after the other. Find the probability that

(i) First two are boys and third is sert.

(ii) First and third of same sex and second is of opposite sex.

Ans : There are 10 boys and 5 girls. So that 15.

(i) The probability that first one is a boy is 10/15

The probability that the second is a boy if first is a boy is $\frac{10}{15}.\frac{9}{14}.$

The probability that first two are boys and third is a girl is

$$\frac{10}{15}.\frac{9}{14}.\frac{5}{13}=\frac{15}{91}$$

(ii) There are two possibilities first, third are boys and second is a girl.
Another possibility is first , third are girl and second is a boy.
The required probability is the sum of these two

$$\frac{10}{15}.\frac{5}{14}.\frac{9}{13}+\frac{5}{15}.\frac{10}{14}.\frac{4}{13}=\frac{5}{21}$$

13. A problem in statistics is seven to three students A, B, C whose chance of solving it are ½,3/4,and ¼. Respectively that is the probability that the problem is solved.

Ans : P(A)= The probability that A to solve that problem.

P(B) = The probability that B to solve the problem.

P(C) = The probability that C to solve the problem.

P(A) = ½, P (B) = ¾, P(C) = ¼.

The required probability = $P(A\cup B\cup C)=1-P(A\cup B\cup C)^1$

= 1-p ($(A^1\cap B^1\cap C^1)$

= 1- $P(A^1).p(B^1).p(c^1)$

A,B,C, are independent events so A^1,B^1,C^1 are independent.

2.1 Introduction:

In the theory of probability is we consider the probability of occurrence of more than one event in succession some times the sequences of order in when the events occur makes a differences and some times it will not make any difference. For example from a box

containing ‘9’ cards of identical size marked with the digits 1,2,3,4,5,6,7,8,9, let us draw tow cards one after the other. This is suggested in two ways.

(i) The card drawn in the 1st draw is placed back into the box before the second draw.

(ii) The card drawn in the 1st draw is not placed back into the box before the second draw.

According to the first scheme the probability of drawing ‘9’ in the second draw will be the same what ever may be the result of the first draw. Where as according to the second scheme probability of drawing 9 in the second draw depends on the result of the first draw. The second scheme gives rise to the notion of the conditional probabilities . In this lesson we discuss the need for conditional probability, its definition, independent events, applications, the Bayes theorem its importance in evaluating probabilities .

2.2 Conditional probability:

Let us consider two events A and B where P(A/B) is the conditional probability of happening of A, given that B has already happened. It is defined as

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

For the above definition to be valid p(B) not equal to 0. Similarly the conditional probability of happening of B when the event A has already happened is denoted by P(B/A) and is defined as .

$$p(B/A) = \frac{P(B \cap A)}{P(A)} WHEREP(A) \neq 0$$

Since $A \cap B$ is same as $B \cap A$ we can write that

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

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

Example:

A bag contains 10 gold and 8 silver coins. Two successive draws of one coin tin each draw are made such that the coin drawn in the first draw is not replaced before the second draw is made . Find the probability that both the draw gives gold coins.

Ans:

Let A, B be the events of drawing a gold coin in the first draw and second draw respectively we are to find $p(A \cap B) = P(A).P(B/A)$.

We know that, $p(A) = \frac{10}{18}$

P(B/A)= Probability of drawing a gold coin in the second draw given that a gold coin is drawn in the first draw. Since the com drawn is not replaced we will have a total of 17 coins of which 9 could be gold and hence the probability of drawing a gold coin in the second draw given that a gold coin is drawn in the first drawn =9/17

i.e p(B/A)=9/17

$$P(A \cap B) = P(A).P(B/A) = \frac{10}{18}.\frac{9}{17} = \frac{5}{17}$$

2.3 Multiplication Rule:

Let A be any two events such that P(A) not equal to 0, P(B) is not equal to 0. Then by definition we know that

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

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

By cross multiplication we get

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

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

These two equations are called multiplication rule of probability for two events we can establish multiplication rule of probability for n events.

2.3.1 Theorem": A_1, A_2,……A_n are events. Then $P(A_1 \cap A_2 ..... \cap A_n) = P(A_1).P(A_3/A_1 \cap A_2).....P(A_n/A_1 \cap A_2 \cap ..... \cap A_{n-1})$

Proof: We prove the result by the principle of mathematical induction. It is obvious that the minimum number of events for the definition of condition number of events for the definition of condition probability is two therefore in our theorem n>=2. For two events the statement of the theorem is

$p(A_1 \cap A_2) = P(A_1).P(A_2/A_1)$ and this follows from the definition and cross multiplication of conditional probability

Let n=3 then L.H.S is

$P(A_1 \cap A_2 \cap A_3) = P(A_1 \cap B) WHERE, B = A_2 \cap A_3$

$P(A_1 \cap B) = P(A_1).P(B/A_1)$ Since the statement is time for two events.

i.e $P(A_1 \cap A_2 \cap A_3) = P(A_1)[((A_2 \cap A_3)/A_1)]$

$$= P(A_1).\left[\frac{P(A_2 \cap A_3 \cap A_1}{P(A_1)}\right]$$

$$= P(A_1)\left[\frac{P(A_2 \cap A_3 \cap A_1}{P(A_1 \cap A_2)} . \frac{P(A_1 \cap A_2}{P(A_1)}\right]$$

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

Hence the results is proved for 3 events in a similar manner suppose the result is the true for n=k

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

WE shall prove that

$P(A_1 \cap A_2 \cap A_3 \cap ..... \cap A_{K+1})$

$= P(B \cap A_{K+1}) where, (B = A_1 \cap A_2 \cap .....A_K)$

$P(B \cap A_{K+1}) = P(B).P(A_{K+1}/B)$

$= P(A_1 \cap A_2 \cap A_3 \cap ..... \cap A_K)P(A_{K+1}/A_1 \cap A_2 \cap .....A_K)$

Since we have assumed that the result is three for n=k the above becomes.

$= P(A_1).P(A_2/A_1).P(A_3/A_1 \cap A_2).....P(A_K/A_1 \cap A_2 \cap ..... \cap A_{K-1}).P(A_{K+1}/A_1 \cap A_2 ..... \cap A_K$

Therefore the result is true for any natural number n by the principal of mathematical induction.

2.4 Independence of Events:

Definition: Two events A,B are said to be statistically independent if the probability of their joint occurrence is same as the product of the probabilities of their individual occurrences.

Symbolically $p(A \cap B) = P(A).P(B)$

2.4.2 Definition: In the case of three events A_1, A_2, A_3 the concept of independence is of two types pairurse independent and mutual independence .if the events are independent taken two at a time we say that they ae pairwise independent.

In the case of three events this menas.

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

In addition to this if the events are independent taken all at a time (in the case of three events).

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

In all the above four conditions are true we say that a set of three events A1, A2, A3, are mutually independent.

In general if we have n events say A_1, A_2, ……..A_n

We say that these are pairwise independent Y

$$P(A_i \cap A_j) = p(A_i)\,p(A_j)\ forall, i \neq j, i, j = 1,2,3,....n$$

These conditions are n_{c_2} in number these in addition to the above n_{c_2} conditions the following are also true.

$$P(A_i \cap A_j \cap A_K) = P(A_i)P(A_j)P(A_K)P(A_k).p(A_l).....i \neq 3 \neq k \neq l.$$

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

These sets of conditions are respectively n_{c_2} , n_{c_3} ……. n_{c_n} is in number.

Hence total number of conditions required for the mutual independence of n events is

$$n_{c_2} + n_{c_3} + ... + n_{c_n} = (1+1)^n - n_{c_1} - n_{c_n} = 2^n - n - 1$$

On the other hand number of conditions required for pairwise independence of n events is only n_{c_2} . This is true for n>=2.

It can be seen that mutually independent events are always pairwise independent while the comerse is not three as can be explained by the following example.

2.4.3 example: Consider box containing 4 cards marked with the digits 100,010,001,111. If A,B,C, be the events representing drawing a card at random with be in hundredth place, one in 10^{th} place , one in 1^{st} place respectively. Then it can be seen that.

$$P(A) = \frac{2}{4} = \frac{1}{2}, P(B) = \frac{1}{2}, P(C) = \frac{1}{2}$$

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

$$P(B \cap C) = \frac{1}{4} P(B).P(C) = \frac{1}{2}.\frac{1}{2} = \frac{1}{4}$$

$$P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = P(A).P(B).P(C)$$

Hence the events A,B,C are pair wise independent but not mutually independent.

2.4.4 Example: if A and B are independent then and A and $\overline{B}$, $\overline{A}$ and $\overline{B}$ are also independent where , $\overline{A}$, $\overline{B}$ are complimentary events A, B, respectively

SOL: Given $P(A \cap B) = P(A).P(B)$

To show that $P(A \cap \bar{B}) = P(A).P(\bar{B})$

We know that $P(\bar{B}) = 1 - P(B)$

Multiplying with p(A) we get

P(A). $P(\bar{B})$ =P(A)-P(A).P(B)

$P(A) - P(A \cap B). - - - - - - - - - - -(1)$

$\because A, B,$ are independent

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

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

$= P(A \cap \bar{B})$ using properties of sets in a similar way we can prove that

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

2.4.5 Example : Given that $P(A_1 \cup A_2) = \frac{5}{6}, P(A_1 \cap A_2) = \frac{1}{3}, P(\bar{A_2}) = \frac{1}{2} FIND, P(A_1), P(A_2)$.
Hence show that A1,A2 are independent.

Solution : $P(\bar{A}_2) = \frac{1}{2} \Rightarrow P(A_2) = \frac{1}{2}$

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

$$\frac{5}{6} = P(A_1) + \frac{1}{2} - \frac{1}{3}$$

$$\Rightarrow P(A_1) = \frac{5}{6} - \frac{1}{2} + \frac{1}{3} = \frac{4}{6} = \frac{2}{3}$$

$$P(A_1).P(A_2) = \frac{2}{3}.\frac{1}{2} = \frac{2}{6} = \frac{1}{3}$$

We know that $P(A_1 \cap A_2) = \frac{1}{3}$

Hence $P(A_1 \cap A_2) = P(A_1)P(A_2)$
$\therefore A_1, A_2$

Are independent.

2.5 Bayes Theorem :

Let $E_1, E_2, \ldots\ldots E_n$ be mutually disjoint exhaustive events with $P(E_i) \neq 0$. IF A is any even that can occur with any of $E_1, E_2, \ldots\ldots E_n$ then

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

The L.H.S is called posterior or inverse probability in the numerator of R.H.S namely $P(E_i)$ is prior probability.

Proof: Given that $\overset{n}{\underset{i=1}{U}} E_i = \Omega$ the sample space and $E_i \cap E_j \neq \Phi, for, i \neq j, i \neq 3 = 1,2,....n$

$$A = A \cap \Omega = A \cap (\overset{n}{\underset{i=1}{U}} E_i)_n$$

$$P(A_i) = \sum_{i=1}^{n} (A \cap E_i) = \sum$$

Consider $P(A \cap E_i) = P(A).P(E_i / A)$

$$\therefore P(E_i / A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(E_i).P(A/E_i)}{\sum_{i=1}^{n} P(E_i).P(A/E_i)}$$

$$\therefore P(E_i / A) = \frac{P(E_i).P(A/E_i)}{\sum_{i=1}^{n} P(E_i).P(A/E_i)}$$

Hence the theorem is proved.

2.6 Examples:

If the probability that a communication system will have high fidelity is 0.81and the probability that it will have fidelity and selectivity is 0.18 what is the probability that a system with high fidelity will also have high selectivity.

Solution : Let the event B will be the communication system to have high fidelity.

P(B)=0.81 (given)

Let the event to have high fidelity and selectivity is $A \cap B$

$P(A \cap B) = 0.81$

By definition of conditional probability we have

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

Probability to have high selectivity if it will have fidelity = p(A/B)=0.18/0.81=2/9.

Example 2: A can hit a target once in five shots, B. can hit two targets in 3 shots, C can hit once target in 4 shots .What is the probability that 2 shots hit the target ?

Solution : Let p(A) be the probability that A hit the target =1/5.

P(B) be the probability that B hit the target =2/3.

P© be the probability that c hit the target = ¼.

The probability that two shots hit the target.

$$= P\left[A \cap B \cap C^1) \cup (A \cap B^1 \cap C) \cup (A^1 \cap B \cap C)\right]$$

These are independent and mutually exclusively also

$$= P\left[A \cap B \cap C^1) \cup (A \cap B^1 \cap C) \cup (A^1 \cap B \cap C)\right]$$

$$= P\left[A \cap B \cap C^1) + (A \cap B^1 \cap C) + (A^1 \cap B \cap C)\right]$$

=P(A).P(B).P(C^1)+P(A).P(B^1).P(C)+P(A^1).P(B).P(C)

=P(A).P(B).(1-P(C))+P(A).(1-P(B)).P(C)+(1-P(A)).P(B).P(C)

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

$$= \frac{15}{60}$$

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

2.6.3 Example: A box I contains 5 red balls , 3 white balls box II contains 3 red balls 6 white balls. A box is chosen at random and a ball is drawn and put it into other box. A ball is drawn from second box. Find the probability that both balls are of same colour.

Solution: To select one box probability is =1/2.

Suppose box I is selected and there are 5 red and 3 white

P(R) =5/8, P(w)=3/8.

Suppose Red is selected from box I and put it in box II then P(R) =3/9 and P(W) =6/9 in box II.

Now a ball is selected from box II then P(R) =4/10, P(W) = 6/10.

There are 3 red balls in box II for which one more added from box I . Similarly box II is selected and put it in box I, P(R) =6/9 and P (W)= 3/9. Suppose if white is selected and put it in box I we have p (R) =4/9. Hence there are four paths reading to the same color.

Required probability = $\frac{1}{2}.\frac{5}{8}.\frac{4}{10}+\frac{1}{2}.\frac{3}{8}.\frac{3}{10}+\frac{1}{2}.\frac{3}{9}.\frac{6}{9}+\frac{1}{2}.\frac{6}{9}.\frac{4}{9}=\frac{2227}{4320}$

2.6.4 Example : What is the probability of getting two queens. If we draw two cards from a pack of 52 cards if

(i) with replacement

(ii) without replacement.

Solution : (i) With replacement (The events are independent)

Probability of drawing one queen is 4/52.

Again the card is replace and drawn again

probability of second drawing is 4/52.

Required probability =4/52.4/52=1/169

(iii) Probability of 1st draw is 4/52. And there remains 51 cards and the events are not independent.
Probability of second drawing =3/51
Required probability = 4/52.3/57=1/221.

2.6.5 Example : if probability of an event A is 0.2.

Probability of an event B is 0.3 and probability of $A\cap B=0.08$ are the events independent

Solution : Given P(A)=0.2, p(B)=0.3, $p(A\cap B)=0.08$

Since we know that when two events are independent then

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

$0.08 \neq 0.2.0.3$ (the events are not independent.)

$P(A \cap B) \neq P(A).P(B).$

2.6.6 Example : In a bolt factory machines A, B,C manufacture 20%,30% 50% of the total of their output and 6%, 3%,and 2% are defective. A bolt is drawn at random and found to be defective. What are the probabilities that it is manufactured by machines A, B, and C ?

Solution :

The probability that the bolt was manufactured by machine A is p(A)=20/100=1/5.

Probability that the bolt was manufactured by machine B is P(B)=30/100=3/10.

The probability that the bolt was manufactured by machine c is P©=50/100=1/2.

Suppose the probability that the bolt drawn is defective is P(D).then probability that a defective bolt is drawn from the bolts manufactured by A is P(D/A)=6/600=0.06

Similarly probability that the defective bolt is from machine B is p(D/B)=3/100=0.03.

Also probability that the defective bolt is from machine C is p(D/C)=2/100.=0.02.

Hence the probability that the bolt which is defective manufactured from a is P(A/D). then by Bayes theorem.

$$P(A/D) = \frac{P(A).P(D/A)}{P(A).P(D/A) + P(B).P(D/B) + P(C).P(D/C)}$$

$$= \frac{\frac{1}{5}.0.06}{\frac{1}{5}.\frac{6}{100} + \frac{3}{10}.\frac{3}{100} + \frac{1}{2}.\frac{2}{100}} = \frac{12}{31}$$

$$P(B/D) = \frac{\frac{3}{10}\frac{3}{100}}{\frac{1}{5}.\frac{6}{100} + \frac{3}{10}.\frac{3}{100} + \frac{1}{2}.\frac{2}{100}} = \frac{9}{31}$$

$$\text{P(C/D)} = \frac{\frac{1}{2}.\frac{2}{100}}{\frac{1}{5}.\frac{6}{100} + \frac{3}{10}.\frac{3}{100} + \frac{1}{2}.\frac{2}{10}} = \frac{10}{31}$$

2.6.7 Example : There are three boxes I, II, III, Box I contains 4 red , 5 blue and 6 black balls . Box II contains 3 red, 4 blue and 5 black balls. Box III contains 5 red, 10 blue and 5 black balls. One box is chosen and one ball is drawn from it. What is the probability that

(i) Red ball is drawn

(ii) Blue ball is drawn

(iv) White ball is drawn

Solution : Let E_i be the event that the box is chosen i=1,2,3

Let A be the event that a red ball is chosen

$P(A/E_i)$ = probability that red ball is chosen from the box i=1,2,3.

$P(A/E_1)$=4/15 (there are total of 15 balls in box I)

$P(A/E_2)$=3/12,(there are total of 12 balls in box II)

$P(A/E_3)$=5/20, (There are total of 20 balls in box III)

(i) $P(A)=P(E_1).P(A/E_1)+P(E_2).P(A/E_2)+P(E_3).P(A/E_3)$
BUT $P(E_1)=P(E_2)=P(E_3)=1/3$

$$\therefore P(A)=\frac{1}{3}.\frac{4}{15}+\frac{1}{3}.\frac{1}{4}+\frac{1}{3}.\frac{1}{4}=\frac{23}{90}$$

(II) Let B be the event that blue ball is drawn

P(B)= probability that the blue ball is drawn

$P(B/E_i)$= probability that the blue ball is chosen from the box i=1,2,3.

$$P(B/E_1)=\frac{5}{15}, P(B/E_2)=\frac{4}{12}, P(B/E_3)=\frac{10}{20}$$

$$P(B)=P(E_1).P(B/E_1)+P(E_2).P(B/E_2)+P(E_3).P(B/E_3)$$

$$=\frac{1}{3}\left[\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right]=\frac{7}{18}$$

(ii) Let C be the event that black ball is drawn i.e, p©=probability that the white ball is drawn

$$P(C/E_1)=\frac{6}{15}, P(C/E_2)=\frac{5}{12}, P(C/E_3)=\frac{5}{20}$$

$$P(C)=P(E_1)P(C/E_1)+P(E_2).P(C/E_2)+P(E_3).P(C/E_3)$$

$$=\frac{1}{3}\left[\frac{2}{5}+\frac{5}{12}+\frac{1}{4}\right]=\frac{16}{45}$$

2.6.8 Example : A box I contains eleven cards numbered 1 to 11, box II contains seven cards numbered 1 to 7. A box is selected at random and a card is drawn if the number is even. Find the probability that the card is from box I.

Solution:

$P(E_1)$= probability that box I is chosen.

$P(E_2)$= Probability that box II is chosen.

P(E) = probability that the card is even.

$P(E_1/E)$= probability that number is even it is selected from box I.

$$P(E_1/E) = \frac{P(E_1).P(E/E_1)}{P(E_1).P(E/E_1) + P(E_2).P(E/E_2)}$$

$$P(E_1) = P(E_2) = \frac{1}{2}$$

$$P(E/E_1) = \frac{5}{11}, P(E/E_2) = \frac{3}{7}$$

Since there are 5 even in box I and 3 even in Box II.

$$\therefore P(E_1/E) = \frac{\frac{1}{2}.\frac{1}{5}}{\frac{1}{2}.\frac{5}{11} + \frac{1}{2}.\frac{3}{7}} = \frac{35}{68}$$

2.6.9 Example : In a class 2% of boys and 3% of girls are having blue eyes. There are 30% girls in the class if a student is selected and having blue eyes. What is the probability that the student is a girl ?

Solution let P(A)= probability that a student having the blue eyes.

$P(E_1)$= The student is a girls.

$P(E_2)$ = The student is a boy.

$P(E_1)$ = 0.3, P(E2) = 7

$P(E_1/A)$= IF student selected is having blue eyes she is a girl.

$$P(E_1/A) = \frac{P(E_1).P(A/E_1)}{P(E_1)P(A/E_1) + P(E_2).P(A/E_2)}$$

$$P(A/E_1) = \frac{3}{100}, P(A/E_2) = \frac{2}{100}$$

$$\therefore P(E_1/A) = \frac{\frac{3}{10}.\frac{3}{100}}{\frac{3}{10}.\frac{3}{100} + \frac{7}{10}.\frac{2}{100}} = \frac{9}{23}$$

2.6.10. Example : A business man goes to hotels X, Y,Z. 20%,50%,and 30% of the time respectively . if is known that 5%, 4%, 8% of the rooms in X, Y,Z hotels have faulty plumbring. What is the probability that business man room having faulty plumbring is assigned to hotel Z.

Solution : The probability that the business man goes to hotel X is p(x) =0.2/

The probability that the business man goes to hotel Y is p(Y)=0.5.

And the probability that the business man goes to hotel 2 is p(Z) =0.3.

Suppose A is the event of faulty plumbring

P(A/x)=0.05, P(A/Y)=0.04, P(A/Z)=0.08

Probability that faulty plumbring is assigned to hotel Z is p(Z/A) given by

$$P(Z/A) = \frac{P(Z).P(A/Z)}{P(X).P(A/X) + P(Y).P(A/Y) + P(Z).P(A/Z)}$$

$$= \frac{0.3X\,0.08}{0.2X\,0.05 + 0.5X\,0.04 + 0.3X\,0.08} = \frac{4}{9}$$

2.6.11 Example : These students ABC are in a summing race. A and B have the same probability of winning and each is twice as likely to win as c. Find the probability that B or c wins.

Solution : Assume that the probability of winning A is 2X the probability of winning B is 2x and the probability of winning C is X.

Probability of winning B or C is p(BuC)

$P(B \cup C) = P(B) + P(C) - P(B \cap C)$ (By addition theorem on probability)

$P(B \cap C) = P(B).P(C) = 2X.X = 2X^2$

B and C are independent.

$\therefore 2X + 2X + X = 1 \Rightarrow X = \frac{1}{5}$

Hence $P(B) = \frac{2}{5}, P(C) = \frac{1}{5}$

$\Rightarrow P(B \cup C) = \frac{2}{5} + \frac{1}{5} - \frac{2}{25} = \frac{13}{25}$

2.6.12 Example : The students in a class are selected at random one after the other for an examination.

Find the probability that the boys and girls are alternate if there are 5 boys and 4 girls , 4boys and 4 girls.

Solution: 5 boys and 4 girls and if they are alternative then it is written as

BGBGBGBGB

$$\frac{5}{9}.\frac{4}{8}.\frac{4}{7}.\frac{3}{6}.\frac{3}{5}.\frac{2}{4}.\frac{2}{3}.\frac{1}{2} = \frac{1}{126}$$

4 boys and 4 girls and if they were alternative then it is given by

BGBGBGBGBG

$$\frac{4}{8}.\frac{4}{7}.\frac{3}{6}.\frac{3}{5}.\frac{2}{4}.\frac{2}{3}.\frac{1}{2} = \frac{1}{70}$$

GBGBGBGB=1

Probability $= \frac{1}{70} + \frac{1}{70} = \frac{2}{70} = \frac{1}{35}$

2.6.13. Example : Box A contains 5 red and 3 white marbles and Box B contains 2 red and 6 write marbles. If a marble is drawn from each box, what is the probability that they are both of the same color.?

Solution : Since box A contains 5 red and 3 white then, there are total of 8 marbles and Box B contains 2 red and 6 white marbles therefore there are total of 8 marbles if a marble is drawn from each box, then the probability that they are both or the Red color is

$$P(R) = \frac{1}{2}.\frac{5}{8} + \frac{1}{2}.\frac{2}{8} = \frac{7}{16}$$

Similarly if they are both of the white color the probability is

$$P(W) = \frac{1}{2}.\frac{3}{8} + \frac{1}{2}.\frac{6}{8} = \frac{9}{16}$$

3.1 **Notion of Random Variable:**

When a statistical experiment is performed, all the possible outcomes of it will generate a set called the sample space, denoted by S. Here we are often much interested in its numerical description rather than its specific outcomes. For example consider random experiment of throwing a die. Then 'X' the number of points on the die is a random variable, since 'X' takes the value 1,2,3,4,5,6. Here the random variable 'X' takes the values 1,2,3,4,5 and 6 each with the probability 1/6.

Note that "twice the number of points on a die" which takes the values 2,4,6,8,10,12 is also a random variable and the 'square of number of points on a die" which takes the values 1,4,9,16,25,36 is also a random variable.

Hence a real variable 'X' whose value is determined by the outcome of a random experiment is called a random variable. Thus to each outcome 'W' of a random experiment there corresponds a real number X(w) which is defined for each point of the sample space S.

For example, if a coin tossed, then sample space is

S= {H, T} I.e. S= {w1, w2}, w1= H, W2 = T

Now X (W) = {1, if w=H

{0, if w= T

Here X (w) is a random variable which takes only two values.

3.2 **Distribution Function and its Properties:**

Let x be a random variable then the function F(x) defined for all real x,

$F(x)= P(X \leq x)= P\{\omega : X(\omega) \leq x\}, \quad -\infty < x < \infty$

Is called the distribution function (df) of x.

Properties of Distribution Function:

Property 1: If F(x0 is the distribution function of a random variable x, and y x<y then (a) 0≤F(x) ≤1∀ $x \in R$, F is bounded (b) F(x)≤F9y), F is monotonically non-decreasing.

Proof: (a) Since probability is a non-negative quantity and lies between 0 and 1, i.e, 0≤δ≤1 therefore we can write 0≤P (X≤x) ≤1

=> 0≤ F (x) ≤1 (P(X≤x)=F(x)).

(b) Since F(x) is a monotonically non-decreasing function of x and x,y be any value in R, such that x<y. Also as (-∞,x] is a subset of (-∞,y] we can write

(X,y] = (-∞, y] – (-∞, x]

P (x, y] = P (-∞, y] – P (-∞, x]= P (Y≤y) – P (X≤x)

P (x, y] = F(y) –F(x) --------- (1)

Also since P (x, y] ≥ 0 we have

F (y) – F(x) ≥ 0 (from (1))

=> F(x) ≤ F(y).

Property 2: If F(x) is distribution Function of random variable X, then

(a) $\lim_{x\to-\infty} F(x) = F(-\infty) = 0$ (b) $\lim_{x\to\infty} F(x) = F(+\infty) = 1$

Proof: (a) Let us define that the sequence of events A_n = {x≤n}. Here the sequence {A_n}is a decreasing sequence of events with

$\lim_{n\to\infty} A_n = \emptyset$ --------- (1)

Therefore by the continuity axiom on probability we have

$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = P(\emptyset) = 0$ (From (1)) ----- (2)

But P (A_n) = P (X≤-n) = F(-n) (by definition of distribution fuction)

$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} F(-n) = 0$ (From (2))

i.e, F(-∞) = 0.

(b) Similarly define that the sequences of events $A_n = \{X \le n\}$, here the sequence $\{A_n\}$, is a increasing sequence of events with

$A_n = S$.....(1)

Hence by the continuity axiom on probability we have

$P(A_n) = p(\lim_{n\to\infty} A_n) = P(S) = 1(\because from 1.....(2)$

$\therefore p(A_n) = p(X \le n) = F(n)(by definition of distribution function)$

$\Rightarrow \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} F(n) = 1(\because from (2))$

$i.e F(+\infty) = 0$

Property 3: if F(x) is the distribution function of the random variable X and if X< y then $p(x < X \le y) = F(y) - F(x)$.

Proof: Since the events $x < X \le y$ and $X \le x$, are disjoint and their union is the event $X \le y$. Hence by addition theorem of probability.

$$P(x < X \le y) + P(X \le x) = P(X \le y)$$

$$P(x < X \le y) = P(X \le y) - P(X \le x)$$

$$(\because P(X \le x) = F(x) \,\&\, P(X \le y) = F(y))$$
$$we have P(x < X \le y) = F(y) - F(x)$$

Property4: F(x) is continuous from the right i.e., f(x+0)=F(x) for each x.

Proof: Let $A_n = (x \le x + \frac{1}{n})$ be a sequence of evnents and for a fixed value of x, sequence of events and for a fixed value of x sequence $\{A_n\}$

Is a decreasing sequence of events with

$$\lim_{n\to\infty} P(A_n) = P(X \le x) = F(x)$$

$$or, \lim_{n\to\infty} P(X \le x + \frac{1}{n}) = F(x) (\because A_n = (X \le x + \frac{1}{n}))$$

$$\lim_{n\to\infty} F(x + \frac{1}{n}) = F(x)$$

$$F(x + \frac{1}{\infty}) = F(x)$$

$$or F(X + 0) = F(x).$$

3.3 Discrete Random variable:

If the sample space S contains a finite number of points or count ably infinite number of points, it is called a discrete sample space. A random variable X defined over a discrete sample space is called a discrete random variable.

For example if we collect data about number of persons in families of certain town, then it is certain that number of persons in each family would be in whole numbers. Therefore there.Would be no family with 2.5 or 2.67 or 1.97 persons.

The variable i.e., the number of persons in a family in this case is a discrete random variable. Some more examples of discrete random variable are given below.

Example 1: The number of heads in tossing of a coin.

Example 2: The number of accidents occurred in a year.

Example 3: The number of printing mistakes in each page of a book.

Example 4: The number of telephone calls received by the telephone operator

Example 5: The number of insects survived when an insecticide is sprayed.

3.4 Probability Mass Function:

If 'x' is discrete random variable defined on the sample space s which takes the values x_1, x_2, with each possible outcome X_i, then a number is associated that is $p_i = P(X=x_i) = P(x_i)$, called the probability of x_i, the numbers $P(x_i)$, I = 1, 2, ---- must satisfy the following conditions.

(a) $P(x_i) \geq 0 \; \forall \; i$ (b) $\sum_{i=1}^{\infty} P(x_i) = 1$

The function is called the probability mass function of the random variable x and the set $\{X_i, P(X_i)\}$ is called the probability distribution of the random variable x.

3.5 Continuous Random variable:

If sample space S contain an infinite number of points or continuity of

Points on a line segment or more than one interval of points is called

Continuous sample space. A random variable defined over the

Continuous sample space is called a continuous random variable.

For Example 1: The weight of middle aged people in India lying between

40 kg and 150 kg is a continuous variable

i.e., x(x) = $\{x : 40 \leq x \leq 150\}$

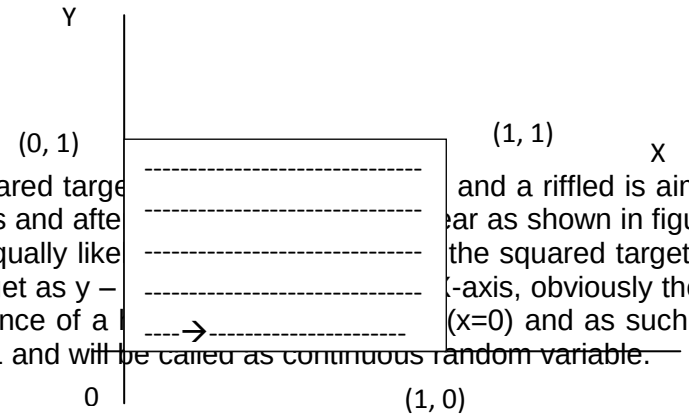
Example 2: The maximum breaking strength 250 kg of a coin is a continuous variable.

i.e., X(x) = $\{x : 0 \leq x \leq 250\}$

Example 3: The amount of rainfall on a rainy day is a continuous variable.

3.6 Probability Density Function.

Since in dealing with the distribution of a continuous random variable will be necessary to express the probabilities in the form of intervals. So with the help of an example continuous probability distribution is explained as follows.



Let us consider a squared targe and a riffled is aimed at it which is triggered several times and afte ear as shown in figure 6.6 whenever a bullet is fixed it is equally like the squared target. Let us consider the left side of the target as y – -axis, obviously the variety X would define horizontal distance of a (x=0) and as such it may have any value between 0 and 1 and will be called as continuous random variable.

From geometric probability it can be seen that the Chance a hit into any interval is equal to the horizontal length of that interval divided by the total length of the board which will be equal to one. For instance the probability that the hit strikes between 0.3 and 0.8, horizontal distance 0.5/1.0=0.5. Since for the continuous distribution it is not possible to have a finite probability associated with single point as in the discrete distribution.

The probability density function for the continuous random variable 'x' in the interval (a,b) is given by

$$f(x)=\begin{cases}0, x<a\\ \phi(x), a\le x\le b\\ 0, x>b\end{cases}\Bigg\}$$

Note 1: $f(x)\ge 0, -\infty < x < \infty$

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

3: the probability p(E) is given by p(E)= $\int_{E} f(x)dx$.

Note 4: Transformation of one dimensional Random variable

Let x be a continuous r.v with p.d.f f(x) and y=g(x).

Be strictly monotonically increasing (or decreasing) function of x. Assume that g(x) is differentiable and hence continuous forall x. then the p.d.f of the R.V. y is given by

3.7 Mathematical Expectation:

Let 'X' be a random variable then the mathematical expectation of 'X' is denoted by E(X) and is given by

$E(X)=\sum_{x} xp(x)$ for discrete random variable and p(x) is the probability mass function.

$=\int_{-\infty}^{\infty} xf(x)dx$ for continuous random variable and f(x) is the probability density function.

R th(moment origin)

For the probability distribution f(x), the rth moment about origin defined as

$\mu_r^1=\int_{-\infty}^{\infty} x^r f(x)dx = E(X^2)$ thus

$\mu_1^1 = E(X), \mu_2^1 = E(X^2)$

Mean = $\bar{x} = \mu_1^1 = E(X)$

AND VARIANCE = $\mu_2 = \mu_2^1 - \mu_1^{1^2} = E(X^2) - [E(X)]^2$

The above result gives the variance in terms of expectations now

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

$$= \int_{-\infty}^{\infty} \left\{x-\bar{x}\right\}^r f(x)dx$$

This gives the r[th] moment about mean and it is denoted by μ_r

Thus $\mu_r = \int_{-\infty}^{\infty} (x-\bar{x})^r f(x)dx$

Put r=1, we get

$$\mu_r = \int_{-\infty}^{\infty} (x-\bar{x}) f(x)dx = \int_{-\infty}^{\infty} xf(x)dx - \int_{-\infty}^{\infty} \bar{x} f(x)dx$$

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

=0

Put r=2, we get,

Variance = $\mu_2 = E\left[\{X-E(X)\}\right] = \int_{-\infty}^{\infty} (x-\bar{x})^2 f(x)dx$

Addition theorem of expeclation.

If X and y be continuous random variables with marginal p.d.f $f_x(x), and f_y(y)$ and whose joint p.d.f $f_{xy}(x,y)$.

$$E(X) = \int_{-\infty}^{\infty} x.f_y(x)dx \text{ ---------(1)}$$

$$E(Y) = \int_{-\infty}^{\infty} y.f_y(y)dy \text{ ----------(2)}$$

Now $E(x+y) = \int_{-\infty-\infty}^{\infty\infty} (x+y) f_{xy}(x,y)dxdy$

$$= \int_{-\infty-\infty}^{\infty\infty} xf_{xy}(x,y)dxdy + \int_{-\infty-\infty}^{\infty\infty} yf_{xy}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} [\int_{-\infty}^{\infty} xf_{xy}(x,y)dxdy] + \int_{-\infty}^{\infty} Y[\int_{-\infty}^{\infty} f_{xy}(x,y)dx]dy$$

$$\int_{-\infty}^{\infty} x f_x(x)dx + \int_{-\infty}^{\infty} y f_y(y)dy$$

=E(X)+E(Y)

Using marginal distribution function

$$x,\ f_x(x) = \int_{-\infty}^{\infty} f_{xy}(x,y)dy$$

$$y,\ f_y(y) = \int_{-\infty}^{\infty} f_{xy}(x,y)dx$$

Multiplication Theorem of Expectations

If X and Y are independent variables , then

E(XY)=E(X).E(Y)

Proof : Let X and Y be continuous random variables with joint p.d.f $f_{xy}(x,y)$ and marginal p.d.f's $f_x(x) and f_y(y)$ respectively.

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

$$E(Y) = \int_{-\infty}^{\infty} y.f_y(y)dy$$

Now E(XY)= $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_x(x).f_Y(y)dx.dy (\because X and Y are, independent)$

$$= \int_{-\infty}^{\infty} x.f_X(x)dx$$

$$= \int_{-\infty}^{\infty} y.f_y(y)dy$$

E(XY)=E(X).E(Y)

Note 1: if $x_1,x_2,\ldots..x_n$ are nindependent random variables then ($E(x_1,x_2,\ldots.x_n)$ = $E(x_1).E(X_2)\ldots\ldots..E(X_n)$

2. If X is a random variable and 'a' is a constant then (a) E(aG(X))=a.E(G(X)) (b) E(G(x)+a)=E(G(x)+a

Where G(x) is a function of 'X' which is also a random variable.

3. if X is a random variable and 'a' and 'b' are constants then E(ax+b)=aE(X)+b

If b=0, then we get E(a,x)=E(X)

If a=1, b=-E(X)=-$\bar{x}$ then we get $E(X-\bar{X})=E(X)-E(X)=0$

4. Let $X_1,X_2,\ldots..X_n$ be any 'n' random variable and $C_1,C_2,\ldots..C_n$ are constant , then

$E(a_1X_1+a_2X_2+\ldots\ldots a_nX_n)=a_1E(X_1)+a_1E(X_2)+\ldots..+a_nE(X_n)$

Which is the expectation of a linear combination of random variables

5. if x is a random variable, then

$V(ax+b)=a^2v(x)$

Where a,b are constants

Covariance :

If X andy are random variables, then covariances between then is defined as

Cov(x,y)=E{[X-E(x)][y-E(Y)]}

=E{XY-XE(Y)-E(X)Y+E(X)E(Y)}

=E(XY)-E(X)(Y)-E(X)E(Y)+E(X)E(Y)

COV(X,Y)=E(XY)-E(X).E(Y)

If X and Y are independent then

E(XY)=E(X).E(Y)

THEN cov (X,Y)=0

Note 1.Cov(ax,by)=abcov(x,y)

2. cov(x+a,y+b)=cov(x,y)

3.cov(ax+b,cy+d)=ac cov (x,y)

4. $v(x_1+x_2)=v(x_1)+v(x_2)+2cov(x_1,x_2)$

5. $v(x_1-x_2)=v(x_1)+v(x_2)-2cov(x_1,x_2)$

If x_1,x_2 are independent cov $(x_1,x_2)=0$ then $v(x_1+-x_2)=v(x_1)+v(x_2)$

<u>Moment Generating Function</u>

The moment generating function (m.g.f) of a random variable 'X ' (about origin whose probability function f(x) is given by

$$M_X(t)=E(e^{tx})=\{\int_{-\infty}^{\infty}e^{tx}f(x)dx,\ for continuous probability functi on$$

$$\sum_{x}^{\infty}e^{tx}p(x),\ for discrete probability,\ function$$

To find the rth moment about origin we know that

$$M_X(t) = E(e^{tx}) = E\left[1 + t \times \frac{1}{2!}(tx)^2 + \frac{1}{3!}(tx)^3 + ... + (tx)^2 + ...\right]$$

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

i.e, $M_X(t) = \sum_{r=0}^{\infty} \frac{t^2}{r!}\mu_r^1, u \sin g, \mu_r^1 = E(X^r)]$

This gives the mgfinterms of moments. Thus the coefficient of $\frac{t^r}{0!}$ in $M_x(t) gives$ the r[th] moment about origin μ_r^1

Since $M_x(t) gives$ generates moment, it is known as moment generating function.

Note : 1 The moment generating function of X about the point x=a is given by

$$M_X(t) = E(e^{t(x-a)})$$

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

$$= 1 + tE(x-a) + \frac{t^2}{2!}E(x-a)^2 + ... + \frac{t^2}{r!}E(x-a)^r + ...$$

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

$$M_X(t)_{x=a} = 1 + t\mu_1^1 + \frac{t^2}{2!}\mu_2^1 + ... + \frac{t^2}{r!}\mu_r^1 + ...$$

2. $M_{cx}(t) = E(e^{tcx}) - - - - - - - - (1)$

$$M_x(ct) = E(e^{ctx}) - - - - - - - - - - - - - - - - - (2)$$

From (1) & (2), we get $M_{cx}(t) = M_x(ct)$

3. if $x_1, x_2, ... x_n$ are n independent random variables then $M_{x1+x_2+.....x_n}(t) = M_{x1}(t) m_{x2}(t) ..... m_x(ct)$

Proof: By the definition

$$M_{x1} + x_2 + ...x_n(t) = E(e^{t(x1+x2+....+xn)})$$

$$= E(e^{tx}).E(e^{tx_2}).....E(e^{tx_n})$$

$$= E(e^{tx}).E(e^{tx_2}).....E(e^{tx_n})$$

$\because x_1, x_2, ...x_n$ are independent

$$= M_{x1}(t).M_{x2}(t).....Mx_n(t).$$

4. if $U = \frac{x-a}{n}$ then $M_u(t) = e^{\frac{-at}{n}}.M_x(\frac{t}{n}), a, harecons\tan ts$

By definition $M_u(t) = E(e^{tu})$

$$= E\left[e^t\left(\frac{x-a}{h}\right)\right]$$

$$= E\left[e^{\frac{tx}{h}}e^{\frac{-at}{h}}\right]$$

$$= e^{-at/h}.E(e^{tx/h})$$

$$= e^{-at/h}.M_X(t/h)(\because bydefiniti onofmgf$$

$$M_u(t) = e^{-at/h}.M_X(t/h)$$

3.8 Worked out example:

3.8.1 Example : A random variable 'X' has the following probability function.

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

(i) Determine the value of 'a'

(ii) Find p(x<3), $p(x \geq 3)$, p(0<x<5)

(iii) Find the distribution function of X.

Solution (i) we know that if p(x) is the probability mass function then $\sum_{i=1}^{\infty} p(x_i) = 1$

(But here 'I' varies from 0 to 8)

$$\sum_{i=1}^{\infty} p(x_i) = 1 \Rightarrow a + 3a + 5a + 7a + 9a + 11a + 13a + 15a + 17a = 1$$

$$81a = 1$$

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

(ii) p(x<3)=p(0)+p(1)+p(2)

=a+3a+5a

$= \frac{9}{81}$

$p(x \geq 3) = p(3) + p(4) + p(5) + p(6) + p(7) + p(8)$

=1-p(x<3)

=1-[p(0)+p(1)+p(2)

$= 1 - \frac{9}{81}$

$= \frac{72}{81}$

(iv) P(0<x<5)=p(1)+p(2)+p(3)+p(4)
=3a+5a+7a+9a
$= \frac{3}{81} + \frac{5}{81} + \frac{7}{81} + \frac{9}{81}$
$= \frac{24}{81}$

(v) To find the distribution function F(x)

x F(x)= $p(x \leq x)$

$0, a(\because p(0) = a$

$1, a + 3a = 4a(\because p(x \leq 1) = p(0) + p(1)$

$2, 4a + 5a = 9a(\because p(x \leq 2) = p(0) + p(1) + p(2)$

$3, 9a + 7a = 16a(\because p(x \leq 3) = p(0) + p(1) + p(2) + p(3)$

$4, 16a + 9a = 25a(\because p(x \leq 4) = p(0) + p(1) + ...p(4)$

$5, 25a + 11a = 36a(\because p(x \leq 5) = p(0) + p(1) + ...p(5)$

$6, 36a + 13a = 49a(\because p(x \leq 6) = p(0) + p(1).....p(6)$

$7, 49a + 15a = 64a(\because p(x \leq 7) = p(0) + p(1).....p(7)$

$8, 64a + 17a = 81a(\because p(x \leq 8) = p(0) + p(1).....p(8)$

3.8.2 Example : suppose that the random variable ‘x’ assumes three values 0,1,2 with probabilities $\frac{1}{3}, \frac{1}{6}, and \frac{1}{2}$ respectively . obtain the distribution function of X.

Solution: Given that

X	0	1	2
P(x=x)	1/3	1/6	½

$x, F(x) = p(x \le x)$

$0, F(0) = p(x \le 0) = \frac{1}{3}$

$1, F(1) = p(x \le 1) = p(0) + p(1) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}$

$2, F(2) = p(x \le 2) = p(0) + p(1) + p(2) = \frac{1}{3} + \frac{1}{6} + \frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1$

3.8.3 Example : Two dice are thrown. Let X assign to each point (a,b) in s the maximum of its numbers i.e, x(a,)=max(a,b).

Find the probability distribution, X is a random variable with

X(s)=(1,2,3,4,5,6)

Solution: p(1)=p(x=1)=p((1,1)=1/36

P(2)=p(x=2)=p[(2,1),(2,2),(1,2)]=3/36

P(3)=p(x=3)=p[(1,3),(3,1),(2,3)(3,9),(3,3)=5/36

P(4)=p(x=4)=p[(1,4),(4,1),(2,4),(4,2),(3,4),(4,3),(4,4)]=7/36

P(5)=p(x=5)=p[(1,5),(5,1),(2,5),(5,2),(3,5)(5,3),(4,5),(5,4),(5,5)=9/36

P(6)=p(x=6)=p[(1,6),(6,1),(2,6),(6,2),(3,6),(6,3),(4,6),(6,4),(5,6),(6,5),(6,6)=11/36

Probability distribution is

X	1	2	3	5	6	7
P(x=x)	1/36	3/36	5/36	7/36	9/36	11/36

3.8.4 Example : if $f(x) = e^{-x}, x \ge 0$
$= 0, x < 0$

(i) is the given function defined above is a density function ?

(ii) if so determine the probability that the variate having this density will fall in the interval (1,2)

(iii) Also find the comulative probability function F(2)=?

Solution : (i) Since $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_{0}^{\infty} f(x)dx$$

$$= \int_{-\infty}^{0} e^{-x}dx + \int_{0}^{\infty} dx$$

$= 0 + \left[-e^{-x}\right]_0^{\infty}$

$= -(e^{-\infty} - e^{0})$

= - (-1)

=1

Hence f(x) satisfies the condition for the function to be a density function.

(ii) In the interval (1,2), e^{-x} is alloways positive

I.e $f(x) \geq 0, in(1,2)$

$\therefore\ p(1 \leq x \leq 2) = \int_1^2 f(x)dx$

$= \int_1^2 e^{-x}dx = e^{-x^2}$

$= -e^{-2} + e^{-1} = 0.368 - 0.135 = 0.233$

(iii) comulative probability function:

$F(x) = p(x \leq x) = \int_{-\infty}^{x} f(X)dx$

Then $F(2) = p(x \leq 2) = \int_{-\infty}^{2} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_0^2 f(x)dx$

$= \int_{-\infty}^{0} e^{-x}dx + \int_0^2 e^{-x}dx$

$= 0 + -e^{-x}{}_0^2 = -e^{-2} + 1$

=1-0.135

=0.865

3.8.5 Example : A continuous random variable 'x' has a p.d.f $f(x) = 3x^2, 0 \leq x \leq 1.$ Find a and b such that

Solution : Given $\begin{matrix} f(x) = 3x^2, 0 \leq x \leq 1. \\ 0, otherwise \end{matrix}$

Then (i) $p(x \leq a) = \frac{1}{2} and p(x > a) = \frac{1}{2}$

When $p(x \le a) = \frac{1}{2} \Rightarrow \int_0^a f(x)dx = \frac{1}{2}$

$\Rightarrow \int_0^a 3x^2 dx = \frac{1}{2}$

$\Rightarrow 3\left(\frac{x^3}{3}\right)_0^a = \frac{1}{2}$

$\Rightarrow a^3 = \frac{1}{2} \Rightarrow a = \left(\frac{1}{2}\right)^{1/3}$

(ii) p(x>b)=0.05

$\Rightarrow \int_b^1 f(x)dx = 0.05$

$\Rightarrow \int_0^a 3x^2 dx = \frac{1}{2}$

$\Rightarrow 3.\frac{x^3}{3}|_b^1 = 0.05$

$\Rightarrow 1 - b^3 = \frac{1}{20}$

$\Rightarrow b^3 = 1 - \frac{1}{20} = \frac{19}{20}$

$\Rightarrow b = \left(\frac{19}{20}\right)^{\frac{1}{3}}$

3.8.6 Example : The diameter of an electric cable, say X assumed to be a continuous random variable with p.d.f $f(x) = 6x(1-x), 0 \le x \le 1.$ (i) Check that the above is a p.d.f (ii) determine a number 'b' such that p(x<b)=p(x>b).

Solution : (i) In the interval $o \le x \le 1, f(x) is always$ positive

I.e in $0 \le x \le 1, f(x) > o$

Then $\int_0^1 f(x)dx = \int_0^1 6x(1-x)dx = 6\int_0^1 \left(x - x^2\right)dx = 6\left[\int_0^1 xdx - \int_0^1 x^2 dx\right]$

$= 6\left[\frac{x^2}{2}|_0^1 - \frac{x^3}{3}|_0^1\right]$

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

$\therefore$ $f(x)$ is a polt of a random variable X.

(ii) Given p(x<b)=p(x>b)

$$\int_0^b f(x)dx = \int_b^1 f(x)dx \Rightarrow 6\int_0^b \left(x - x^2\right)dx = 6\int_b^1 \left(x - x^2\right)dx$$

$$\Rightarrow \left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^b = \left[\frac{x^2}{2} - \frac{x^3}{3}\right]_b^1 \Rightarrow \left[\frac{b^2}{2} - \frac{b^3}{3}\right] = \left[\frac{1-b^2}{2} - \frac{(1-b^3)}{3}\right]$$

$$\frac{b^2}{2} - \frac{1+b^2}{2} = \frac{b^3}{3} - \frac{(1-b^3)}{3}$$

$$\frac{b^2 - 1 + b^2}{2} = \frac{b^3 - 1 + b^3}{3}$$

$$\frac{2b^2 - 1}{2} = \frac{2b^3 - 1}{3}$$

$$\frac{2b^3 - 1}{3} - \frac{2b^2 + 1}{2} = 0 \Rightarrow \frac{4b^3 - 2 - 6b^2 + 3}{6} = 0$$

$$\Rightarrow 4b^3 - 6b^2 + 1 = 0$$

$$\Rightarrow (2b - 1)(2b^2 - 2b - 1) = 0$$

$$\Rightarrow b = \frac{1}{2}; b = \frac{2 \pm \sqrt{4-8}}{4} = \frac{2 \pm i2}{4} = \frac{1 \pm i}{2}$$

Here b=1/2 is real value and $b = \frac{l \pm i}{2}$ is imaginary. therefore b=1/2 which lies m(0,1).

Note : Let 'X'be a random variable with p.d.f f(x) which is defined in the interval (a,b), then

(i) Arithmetic mean = $\int_a^b xf(x)dx$

(ii) Harmonic Mean = $\int_a^b \frac{1}{x} f(x)dx$

(iii) Geometric mean 'G' is given by logG= $\int_a^b \log xf(x)dx$

(iv) Moments about origin $\mu_r^1 = b\int_a^b x^8 f(x)dx$

(v) moment about any point A $\mu_r^1 = \int_a^b (x-A)^2 f(x)dx$

(vi) Moment about mean $\mu_r = \int_a^b (x-mean)^2 f(x)dx$

(vii) Mean deviation about the mean is

M.D = $\int_a^b |x-mean| f(x)dx$

3.8.7. Example :

The probability density function is $f(x) = k(3x^2-1), -1 \le x \le 2$
$0, ohterwise$

Find the value of K and find the probability $(-1 \le x \le 2)$

Solution : Given the p.d.f $f(x) = \{k(3x^2-1), -1 \le x \le 2\}$
$o, otherwise$

We know that the total probability is unity

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

$\Rightarrow \int_{-\infty}^{-1} f(x)dx + \int_{-1}^{2} f(x)dx + \int_{2}^{\infty} f(x)dx = 1$

$\int_{-1}^{2} k(3x^2-1)dx = 1$

$k\left[x^3 - x\right]_0^2 = k[+8-1-2-(-1)] = 1$

$\Rightarrow 6k = 1$

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

$f(x) = \{\frac{1}{6}(3x^2-1) in -1 \le x \le 2$
$0, otherwise$

then $p(-1 \le x \le 0) = \int_{-1}^{0} f(x)dx = \int_{-1}^{0} \frac{1}{6}(3x^2 - 1)dx$

3.8.8 Example: if $f(x) = e^{-|x|}$ is p.d.f in $-\infty \le x \le \infty$. find the value of k, variance of the random variable and also find probability between 0 and 4

Solution : Given $f(x) = \{ke^{-|x|}, -\infty \le x \le \infty$
$0, otherwise$

We know that the total probability is unity.

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

$$\int_{-\infty}^{\infty} k.e^{-|x|}dx = \int_{-\infty}^{0} k.e^{-|x|}dx + \int_{0}^{\infty} k.e^{-|x|}dx$$

$$\int_{-\infty}^{0} ke^{x}dx + \int_{0}^{\infty} k.e^{-x}dx$$

$(\because -\infty to 0 |x| = -x, and, in 0 to \infty, |x = x|$

$$= ke^{x}\ |_{-\infty}^{0} + k\int_{0}^{\infty} e^{-x}dx = k + k = 2k = 1$$

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

$$f(x) = \frac{1}{2}e^{-|x|}$$

Now mean is given by

$$\int_{-\infty}^{\infty} xf(x)dx = \frac{1}{2}\int_{-\infty}^{\infty} xe^{-|x|}dx = \frac{1}{2}\int_{-\infty}^{0} xe^{x}dx + \frac{1}{2}\int_{0}^{\infty} x.e^{-|x|}dx$$

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

$$= \frac{1}{2}[-1+1] = 0$$

Variance is $\sigma^2 = E(x^2) - [E(x)]^2$

$$E(X^2) = \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-|x|}dx = \frac{1}{2}\int_{-\infty}^{0} x^2 e^{x}dx + \frac{1}{2}\int_{0}^{\infty} x^2 e^{-x}dx$$

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

$$= \frac{1}{2}(2+2) = \frac{4}{2} = 2$$

$$\therefore \sigma^2 = 2 - 0^2 = 2.$$

Hence variance =2

$$p(0 \le x \le 4) = \frac{1}{2}\int_0^4 e^{-|x|}dx = \frac{1}{2}\int_0^4 e^{-|x|}dx\left(\because in, 0to4, |x = x|\right)$$

$$= \frac{1}{2}\left(-e^{-x}\right)_0^4 = \frac{1}{2}\left(1 - e^{-4}\right) = \frac{1}{2}(1 - 0.018) = 0.491$$

The probability between o and 4 is 0.491.

3.8.9 Example:

Probability density function of a random variable x is ½ sinx in $0 \le x \le \Pi$ a

And is 0 otherwise find the mean, mode and median for the distribution and also find the probability between 0 and $\Pi/2$.

Solution : Given $f(x) = \begin{Bmatrix} \frac{1}{2}\sin x, 0 \le x \le \Pi \\ 0, otherwise \end{Bmatrix}$

We have mean $\int_{-\infty}^{\infty} xf(x)dx$

$$= \int_{-\infty}^{0} xf(x)dx + \int_{0}^{\Pi} xf(x)dx + \int_{\Pi}^{\infty} xf(x)dx$$

$$= 0 + \frac{1}{2}\int_0^{\Pi} x\sin xdx + 0 = \frac{1}{2}\left[-\cos x + \sin x\right]_0^{\Pi} = \frac{\Pi}{2}$$

Next mode is defined as the value of x for which f(x) is maximum.

$f^1(x) = 0 and f^{11}(x)$ is negative at the value of x.

$$f^1(x) = \frac{1}{2}\cos x = 0, when x = \frac{\Pi}{2}$$

$$f^{11}(x) = \frac{-1}{2}\sin x;\ f^{11}(x) = -\frac{1}{2} < 0, when X = \frac{\Pi}{2}, \bmod e = \frac{\Pi}{2}$$

Suppose M is the median then $\int_0^M f(x)dx = \int_M^{\Pi} f(x)dx = \frac{1}{2}$

$$\frac{1}{2}\int_0^M \sin x dx = \frac{1}{2} \Rightarrow \frac{1}{2}(1-\cos x)_0^M = \frac{1}{2}(1-\cos M) = \frac{1}{2}$$
$$= \frac{1}{2} - \frac{\cos M}{2} = \frac{1}{2}$$

CosM should be zero, hence $M = \frac{\Pi}{2}$

The above problem has mean =mode =median $= \frac{\Pi}{2}$

3.8.10 Example:

A continuous random variable x has the distribution function.

$$F(x) = \begin{cases} 0, if x \le 1 \\ k(x-1)^4, if, 1 < x \le 3 \\ 1, if, x > 3 \end{cases}$$

Find K and the probability density function of x.

Solution:

Given $F(x) = \begin{cases} 0, if x \le 1 \\ k(x-1)^4, if, 1 < x \le 3 \\ 1, if x > 3 \end{cases}$

We know that $f^{-1}(x) = f(x)$

Then $f(x) = \begin{cases} 0, if, x \le 1 \\ 4k(x-1)^3 if, 1 < x \le 3 \\ 1, if, x > 3 \end{cases}$

As total probability is unity we have

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow i.e \int_0^3 f(x)dx = 1$$

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

$$4k\frac{(x-1)^4}{4}|_0^3 = 1$$

$$\Rightarrow k = \frac{1}{16}$$

$$f(x) = (x-1)^3, if, 1 < x \le 3.$$

3.8.11 Example:

A continuous random variable ‘x’ is distributed over the interval [0,1] with pdf ax^2+bx, where a,b are constants. If the arithmetic mean of ‘x’ is 0.5, find the values of a and b.

Solution:

Let $f(x)= ax^2+bx$

Which is a pdf in [0,1]

i.e $\int_0^1 f(x)dx = 1$

$$\int_0^1 \left(ax^2 + bx\right)dx = 1$$

i.e $$\left(\frac{ax^3}{3} + \frac{bx^2}{2}\right)_0^1 = 1$$

$$a\left(\frac{1}{3}\right) + b\left(\frac{1}{2}\right) = 1$$

$$2a + 3b = 1 - - - - - - - - - - - - - - (1)$$

Now mean $= \int_a^b xf(x)dx = \int_0^1 xf(x)dx$

$$= \int_0^1 x\left(ax^2 + bx\right)dx$$

$$= \left[\frac{ax^4}{4} + \frac{6x^3}{3}\right]_0^1$$

$$= \frac{a}{4} + \frac{b}{3}$$

Given mean =0.5=1/2.

$$\therefore \frac{1}{2} = \frac{a}{4} + \frac{b}{3}$$

$$\frac{1}{2} = \frac{3a + 4b}{12}$$

i.e 3a+4b=6----------(2)

From 1 and 2 we get after solving.

a=-6,b=6.

3.8.12 Example:

Prove that the geometric mean G of the distribution dF(x)=6(2-x)(x-1)dx, $1 \le x \le 2$ is given by 6log(16G)=19.

Solution : Given dF(x)=6(2-x)(x-1)dx

Pdf f(X)= 6(2-x)(x-1)

$$\log G = \int_1^2 \log x.f(x)dx \left[u \sin g \log G = \int_a^b \log xf(x)dx \right]$$

$$= 6\int_1^2 \log x(2-x)(x-1)dx$$

$$= -6\int_1^2 \left(x^2 - 3x + 2\right)\log xdx$$

$$= -6\left[\int_1^2 \log xd\left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\right]$$

$$= -6\left[\left\{\log x\left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\right\}^2 - \int_1^2 \left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\frac{1}{x}dx\right]$$

$$= \left[\log 2\left(\frac{8}{3} - 3\times 2 + 4\right) - \int_1^2 \left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\frac{1}{x}dx\right]$$

$$= \left[\log 2\left(\frac{8}{3} - 3\times 2 + 4\right) - \int_1^2 \left(\frac{x^3}{3} - \frac{3x}{2} + 2\right)dx\right]\left(\because \log 1 = 0\right)$$

$$= -6\left[\log 2\times\frac{2}{3} - \left(\frac{x^3}{9} - \frac{3x^2}{4} + 2x\right)_1^2\right]$$

$$= -6\left[\frac{2}{3}\log 2 - \left(\frac{8}{9} - 3 + 4 - \frac{1}{9} + \frac{3}{4} - 2\right)\right]$$

$$= -4\log 2 + 6\left(\frac{8}{3} - 1 - \frac{1}{3} + \frac{3}{4}\right)$$

$$= -4\log 2 + 6\left(\frac{19}{36}\right)$$

log=-4log2+19/6.

$$\log G + \log 2^4 = \frac{19}{6}\left(\log m^n = n\log m\right)$$

6log(GX16)=19 (logmn=logm+logn)

3.8.13 Example:

A random variable X has the following probability function.

(i) Find the value of k.
(ii) (ii) mean
(iii) Variance
(iv) $p(\geq 3)$
(v) $p(1 < x \leq 5)$

X	1	2	3	4	5	6
P(X=x)	k	3k	5k	7k	9k	11k

Solution: (i) We know that $\sum_{i=1}^{n} p(x_i) = 1$

K+3k+5k+7k+9k+11k=1

=36k=1

K=1/36.

(ii) mean $\mu = \sum_x xp(x) = 1.\frac{1}{36} + 2.\frac{3}{36} + 3.\frac{5}{36} + 4.\frac{7}{36} + 5.\frac{9}{36} + 6.\frac{11}{36}$

$$= \frac{161}{36} = 4.46$$

(iii) Variance $\sigma^2 = \sum_x (x-\mu)^2 p(x) = E(X^2) - \left[E(X)\right]^2$

$$= 1.\frac{1}{36} + 4.\frac{3}{36} + 9.\frac{5}{36} + 16.\frac{7}{36} + 25.\frac{9}{36} + 36.\frac{11}{36} - (4.46)^2$$
$$= 2.08$$

(iv) $p(x \geq 3) = p(x=3) + p(x=4) + p(x=5) + p(x=6)$

$$= \frac{5}{36} + \frac{7}{36} + \frac{9}{36} + \frac{11}{36} = \frac{32}{36} = \frac{8}{9}$$

(v) $p(1 < x \leq 5) = p(x=2) + p(x=3) + p(x=4) + p(x=5)$

$$= \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36} = \frac{24}{36} = \frac{2}{3}$$

3.8.14 Example:

A sample of 4 items is selected at random from a box containing 12 items of which 5 are dejective .find the expected number E of dejective items

Solutions: Since a sample of 4 items is selected at random from a box containing 12 items then the number of exhaustive cases is 12_{c_2} = 495.

Probability that there are zero defective items is as there are 5 are defective out of 12 items.

$$\frac{7_{c_4}}{495} = \frac{35}{495}$$

Probability that there are one defective item= $\frac{7_{c_3}.5_{c_1}}{495} = \frac{175}{495}$

Probability that there are two defective items $\frac{7_{c2}.5_{c2}}{495} = \frac{210}{495}$

Probability that there are three defective items = $\frac{7_{c_1}.5_{c_3}}{495} = \frac{70}{495}$

Probability that there are four defective items = $\frac{5_{c_4}}{495} = \frac{5}{495}$

X	0	1	2	3	4
P(X=x)	$\frac{7}{99}$	$\frac{35}{99}$	$\frac{42}{99}$	$\frac{14}{99}$	$\frac{1}{99}$

Expected number of defective items =mean= $\bar{x} = \sum_x x.p(x)$

$$= 0.\frac{7}{99} + 1.\frac{35}{99} + 2.\frac{42}{99} + 3.\frac{14}{99} + 4.\frac{1}{99}$$
$$= \frac{165}{99}$$

3.8.15 Example:

For the describe probability distribution

X	0	1	2	3	4	5	6	7

F(x) 0 k 2k 2k 3k k^2 $2k^2$ $7k^2+k$

Determine (i) K, (ii) mean (iii) variance, (iv) smallest value of X such that

$$p(x \le x) \ge \frac{1}{2}$$

Solution : (i) We know that $\sum_i p(x_i) = 1$

i.e. $0+k+2k+2k+3k+k^2+2k^2+7k^2+k=1$

$10k^2+9k-1=0$,

K=-1,1/10

$\because\ p(x) \ge 0, k cannot be -1$

$$k = \frac{1}{10}$$

(ii) $\mu = mean = \sum_x xp(x) = 0+1.\frac{1}{10}+2.\frac{2}{10}+3.\frac{2}{10}+4.\frac{3}{10}+5.\frac{2}{100}+6.\frac{1}{10}+7.(\frac{7}{100}+\frac{1}{10})$

$$\frac{366}{100} = 3.66$$

(iii) Variance = $\sigma^2 = \sum_x x^2 p(x) - (mean)^2$

$$= \frac{1}{10}+4.\frac{2}{10}+9.\frac{2}{10}+16.\frac{3}{10}+25.\frac{1}{100}+36.\frac{1}{100}+49.(\frac{7}{100}+\frac{1}{10})-(3.66)^2$$
$$= 37.7$$

(iv) f(0)+f(1)=0.1

f(0)+f(1)+f(2)=0+0.1+0.2=0.3

f(0)+f(1)+f(2)+f(3)=0.3+0.2=0.5

f(0)+f(1)+f(2)+f(3)+f(4)=0.5+0.3=0.8

The smallest value of x such that $p(x \le x) > \frac{1}{2}, is 4, p(x \le 3) = 0.5, p(x \le 4) = 0.8$

3.8.16 Example:

Find the mgf of the random variable whose moments are $\mu_r^1 = (r+1)!2^r$

Solution : we know that the mgf interms of moments is given by

$$M_x(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r^1$$

$$\sum_{r=0}^{\infty}(r+1)!2^r.\frac{t^r}{r!}(\because \mu_r^1=(r+1)!2r$$

$$=\sum_{r=0}^{\infty}\frac{(2t)^r(r+1)!}{r!}$$

i.e., $M_X(t)=1+2(2t)+3(2t)^2+.....$

$=(1-2t)^{-2}(u\sin g(1-x)^{-2}=1+2x+3x^2+....)$

3.8.17 Example: if the moments of a random variable X are defined by $E(X^2)=0.6$, r=1,2,3,…..

Show that p(x=0)=0.4, p(x=1)=0.6, p(x≥ 2)=0

$$M_x(t)=\sum_{r=0}^{\infty}\frac{t^r}{r!}\mu_r^1$$

$$=1+\sum_{r=1}^{\infty}\frac{t^r}{r!}(0.6)(GivenE(x^r)=\mu_r^1=0.6$$

$$=1-0.6+0.6+\sum_{r=1}^{\infty}\frac{t^r}{r!}(0.6)$$

$$=0.4+0.6[1+\sum_{r=1}^{\infty}\frac{t^r}{r!}]$$

$$M_x(t)=0.4+0.6e^t$$

But $M_x(t)=E(e^{tx})=\sum_{r=0}^{\infty}e^{tx}.p(x)$

$$M_x(t)=p(0)+e^t\,p(1)+\sum_{r=2}^{\infty}e^{tx}.p(x)$$

From 1 and 2 , we get

P(0)=0.4, p(1)=0.6

$$\sum_{r=2}^{\infty}e^{tx}\,p(x)=0\Rightarrow p(x)=0,x>2$$

3.8.18 Example:

Find the m.gf of a random variable ‘x’ having the pdf $f(x)=\begin{cases}\frac{1}{3},-1<x<2\\0,otherwise\end{cases}$

Solution: We know that the m.g.f for a continuous random variable 'x' is

$$M_x(t) = \int_{-\infty}^{+\infty} e^{tx} f(x)dx = \int_{-1}^{2} e^{tx}.\frac{1}{3}dx = \frac{1}{3}\left[\frac{e^{tx}}{t}\right]_{-1}^{2}$$

$$M_x(t) = \frac{1}{3}\left[\frac{e^{2t} - e^{-t}}{t}\right]$$

3.8.19 Example:

Find the m.g.f of the random variable 'x' having p.d.f $f(x) = \begin{cases} x, \ for 0 \le x < 1 \\ 2 - x, \ for 1 \le x < 2 \\ 0, otherwise \end{cases}$

Solution: We know that

$$M_x(t) = \int_{0}^{\infty} e^{tx} f(x)dx (here X is a continuous r.v)$$

$$= \int_{0}^{1} e^{tx} f(x)dx + \int_{1}^{2} e^{tx} f(x)dx = \int_{0}^{1} e^{tx}.xdx + \int_{1}^{2} e^{tx}(2-x)dx$$

$$= \left\{x\left(\frac{e^{tx}}{t}\right) - \left(\frac{e^{tx}}{t^2}\right)\right\}_0^1 + \left\{(2-x)\left(\frac{e^{tx}}{t}\right) - (-1)\left(\frac{e^{tx}}{t^2}\right)\right\}_1^2$$

(by using integration by parts)

$$= \frac{e^t}{t} - \frac{e^t}{t^2} + \frac{1}{t^2} + \frac{e^{2t}}{t^2} - \frac{e^t}{t} - \frac{et}{t^2}$$

$$= \frac{e^{2t}}{t^2} + \frac{1}{t^2} - \frac{2e^t}{t^2}$$

$$M_x(t) = \frac{\left(e^t - 1\right)^2}{t^2}$$

Discrete Distributions:

1. Binominal Distribution

Under Certain assumptions we are to find lute required probability i.e., sometimes need to find lute probability of x successes out of n trials .this assumptions are

Assumptions:

1. In each trial have are only two possible in comes called success and failure.
2. The probability of a success is same from each trial.

3. Then one n trial where n is a constant.
4. The n. Trials are independent.

The trials in an expedient satisfying above assumptions are called Bernoulli trials. Let E be an event and probability of happening of an event is called probability of success and probability of non-happening of an event E is called probability of failure. Q. then p+q=1. Since Bernoulli trials is having only two possibilities success and failure then sample space is S= (EUE)

Suppose that are n trails then probability of getting n success and n-x failures is $p^x.q^{n-x}$. Also the number of combinations of x objects selected among n objects is nc_x. Hence the probability of setting success and n-x features among n trails is $nc_xp^x .q^{n-x} =n_{cx}p^x. (1-P)^{n-x}$ (i.e. q=1-p)

Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700 was first published in 1713. Let a random experiment be performed repeatedly an let the occurrence of an event in a trail be called a success and its non-occurrence a failure consider set of n independent Bemoullian trails (being finite) in which the probability p of success in any trail is constant5 for each trial then q = 1-p is the probability of failure in ant trial.

The probability of x success and consequently (n-x) failures in n independent trials in a specified order *(say) SSFSFFFS----FSF where Represents success and F failure according to compound probability theorem.

P (SSFSFFFS ------ FSF) = P(S) P(S) P (F) P(S) P (F) P (F) P (F) P(S) x---- x P (F) P(S) P (F)

= p.p.q.p.q.q.q x ----- x q.p.q

$$= \underbrace{p.p.p.----p}_{x\ factors}\ \underbrace{q.q.q.----q}_{n-x\ factors}$$

But x success in n trials can occur in (x^n) ways and the probability for each of these ways is $p^x.q^{n-x}$. Hence the probability of x success in n trials is given by the addition theorem of probability what so ever the order is by

$$\binom{n}{x} p^x . q^{n-x}$$

The probability distribution on the number of success so obtained is called the binomial probability distribution the probabilities of 0, 1, 2, --------n success is

$$q^n, \binom{n}{1} q^{n-1}.p, \binom{n}{2} q^{n-2}p^2, ----p^n$$

Are the successive terms of the binomial expansion $(q+p)^n$

Definition: A random variable X is said to follow binomial distribution is it assumes only non negative values and its probability mass function is given by

$$P(X=x) = p(x) = \begin{cases} \binom{n}{x} p^x . q^{n-x} ; x = 0,1,2, --n: q = 1 - [\\ 0, \qquad otherwise \end{cases}$$

The two independent constants n and p in the distribution are known as the parameters of the distribution " n" is also known as the degree of the binomial distribution. Binomial distribution is a discrete distribution as X can take only the integral values i.e. 0, 1, 2, ----n. Any visible which follows binomial distribution is known as binomial variety. X ~B(n,p)denotes the random variable X follows binomial distribution with parameters n and p the p(x) in (1) is also sometimes denoted by (x;n,p).

Example: Ten coins are thrown simultaneously. Find the probability of setting at least seven heads.

Solution: The probability of getting a head p=1/2

The probability of getting a tail q=1/2

Then probability of getting at least x heads in a random thrown of 10coins is

$$P(x) = \binom{10}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{10-x} = \binom{10}{x}\left(\frac{1}{2}\right)^{10} \quad x = 0,1,----,10$$

Probability of getting at least seven heads is given by

P (X≤7) = P (7) + P (8) + P (9) + P (10)

$$= \left(\frac{1}{2}\right)^{10}\left\{\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right\}$$

$$= \frac{120+45+10+1}{1024} = \frac{176}{1024}$$

Mean of the Binominal distribution:

Among n trial x are the successes is given by $\binom{n}{x}p^x.q^{n-x}$ x=0, 1, 2, ----n. then mean of binominal distribution is given by

Mean = $\mu_1^1 = E(X) = \sum_{x=0}^{n} x.\binom{n}{x}p^x.q^{n-x} = np\sum_{x=1}^{n}\binom{n-1}{x-1}p^{x-1}q^{n-x}$

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

Variances of the Binominal Distribution:

Variances of the binominal distribution is given by

Variance (x) = (σ^2) = E (X^2) – $[E(X)]^2$ = $\mu_2^1 - (\mu_1^1)^2$

Where $\mu_2^1 = E(x^2) = \sum_{x=0}^{n} x^2\binom{n}{x}p^x.q^{n-x}$

$$= \sum_{x=0}^{n}\{x(x-1)+x\}\frac{n(n-1)}{x(x-1)}\binom{n-2}{x-2}p^x.q^{n-x}$$

$$= n(n-1)p^2\left[\sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}.q^{n-x}\right] + np$$

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

Variance (x) = (σ^2) = n (n-1) p^2+np-n^2p^2 (μ_1^1= np)

Variance (x) = (σ^2) = np (1-p) = npq (q=1-p)

Mode of the binominal distribution:

We have

$$\frac{P(x)}{P(x-1)} = \binom{n}{x} p^x q^{n-x} \Big/ \binom{n}{x-1} p^{x-1} q^{n-x+1}$$

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

$$= \frac{(n-x+1)p}{xp} = \frac{xq + (n-x+1)p - xq}{xq}$$

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

$$= 1 + \frac{(n+1)p - x}{xq} - - - - - \rightarrow (1)$$

Mode is the value of x for which p(x) is maximum here then arises to cases.

Case: when (n+1) p is not an integer

Let (n+1) p=m+f, where m is an integer and f is fractional such that 0<f<1.

subsituting in (1)we get

$\frac{P(x)}{P(x-1)} = 1 + \frac{m-x}{xq}$ ----------→ (2)

From (2) it is obvious that

$$\frac{P(x)}{P(x-1)} > 1 \; for \; x = 0,1,2, - - - m$$

$$\& \frac{P(x)}{P(x-1)} < 1 \; for \; x = m+1, m+2, - - -, n$$

$$=> \frac{P(1)}{P(0)} > 1, \frac{P(2)}{P(1)} > 1, - - - -, \frac{P(m)}{P(m-1)} > 1$$

$$and \; \frac{P(m+1)}{P(m)} < 1, \frac{P(m+2)}{P(m+1)} < 1, - - - - \frac{P(n)}{P(n-1)} < 1,$$

P(0)<P(1)<P(2)< ---- < p(m-1)<p(m)>p(m+1)>p(m+2)>p(m+3) ----- >P(n)

Thus in this case there exist image model value for binominal distribution and it is m, the integral part of (n+1) p

Case: when (n+1) p is an integer

Let (n+1) p=m can integer

Substitution is (1) we get

$\frac{P(x)}{P(x-1)} = 1 + \frac{m-x}{xq}$ ----------→ (3)

From (3) it is obvious that

$$\frac{P(x)}{P(x-1)} \begin{cases} > 1 \; for \; x = 1,2, - - - m-1 \\ = 1 \; for \; x = m \\ < 1 \; for \; x = m+1, m+2, - - - n \end{cases}$$

Then we have

P (0) <P (1) < ---- < P (m-1) = p (m)>p (m+1)>p (m+2) > ------- > p (n)

Hence in this case the distribution is binominal and the two model values are m and m-1.

Moment generating Function of Binominal distribution:

Let X be a variable following binomial distribution, then

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tX} \binom{n}{x} p^x . q^{n-x}$$

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

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

Note: 1 sum of two independent binominal variegates is not a binominal variant.

i.e, $M_{X+Y}(t) = (q_1+p_1e^t)^{n1} . (q_2+p_2e^t)^{n2}$

Which cannot be expressed in the form $(q+pet)^n$

Worked outexample:

1. Example: For a binominal distribution mean =6 and variance is 2.Find the probability of two successes.

Solution: Given mean np=2

For a binominal distribution

q=2/6, p=1-q=1-1/3=2/3

np =6, p=2/3 n=6/p=6/(2/3)=9

The probability of two successes is given by

$$P(X = 2) = \binom{n}{2} p^2 (q^{n-2}) = \binom{9}{2} \left(\frac{2}{3}\right)^2 \left(\frac{2}{6}\right)^{9-2}$$

$$= 36 . \frac{4}{3^9} = \frac{16}{3^7}$$

$$P(X = 2) = \frac{16}{3^7}$$

2. Example : Determine the binominal distribution for which the mean is 4 and variance 3 and find its mode

Solution: Given that mean =4=np

Variance =3=npq

$$q = \frac{npq}{np} = \frac{3}{4}, p = 1 - q = 1 - \frac{3}{4} = \frac{1}{4}, np = 4 => n = \frac{4}{p} = \frac{4}{1/4} = 16.$$

Then Binominal distribution is $\left(\frac{1}{4} + \frac{3}{4}\right)^{16}$

Mode= (n+1) p=17/4.p

(n+1) p is not an integer then integral part of (n+1) p i.e, 4 is the mode.

3. Example: in a binominal distribution mean = 4 and variance is 2. Find the mode of binominal distribution

Solution: Mean = np=4, variance npq=2 are given for binominal distribution

Then $q = \frac{npq}{np} = \frac{2}{4} = \frac{1}{2}\ then\ p = 1 - q = 1 - \frac{1}{2} = \frac{1}{2}$

$$np = 4, p = \frac{1}{2} => n = \frac{4}{p} = \frac{4}{1/2} = 8$$

Mode= (n+1) p = 9.1/2 not an integer.

Mode = 4. The distribution is unimodal.

4. Example: The mean of binominal distribution is 3 and variance is 9/4.Find (i) the value of n (ii) P (x>1) (iii) P (X≤7) (iv) P (1≤x≤6).

Solution: we are given mean =np =3, variance =npq =9/4 for a binominal distribution then

$$(i) q = \frac{npq}{np} = \frac{9/4}{3} = \frac{3}{4}, \qquad p = 1 - q = 1 - \frac{3}{4} = \frac{1}{4}.$$

$$np = 3 => n = \frac{3}{p} = \frac{3}{1/4} = 3{:}4 = 12$$

$$(ii) P(X \geq 1) = 1 - P(X = 0)$$

$$= 1 - \binom{n}{x} p^x q^{n-x} = 1 - \binom{12}{0}\left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{12-0}$$

$$= 1 - \left(3/4\right)^{12} = 1 - 0.03167 = 0.9683$$

(iii) P (X≤7) = P (X=0) + P (X=1) + P(X=2) + P (X=3) + P (X=4 + P (X=5) + P (X=6) + P (X=7)

$$= \left(\frac{3}{4}\right)^{12} + 12.\frac{1}{4}\left(\frac{3}{4}\right)^{11} + \binom{12}{2}\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right)^{10} + \binom{12}{3}\left(\frac{1}{4}\right)^{3}\left(\frac{3}{4}\right)^{9} + \binom{12}{4}\left(\frac{1}{4}\right)^{4}\left(\frac{3}{4}\right)^{8} + \binom{12}{5}\left(\frac{1}{4}\right)^{5}\left(\frac{3}{4}\right)^{7} + \binom{12}{6}\left(\frac{1}{4}\right)^{6}\left(\frac{3}{4}\right)^{6} + \binom{12}{7}\left(\frac{1}{4}\right)^{7}\left(\frac{3}{4}\right)^{5}$$

$$= \frac{3^7}{4^{12}}[243 + 972 + 1782 + 1980 + 1485 + 72 + 28 + 8]$$

$$= \frac{3^7}{4^{12}}(6570) = 0.8564$$

(iv) P (1≤x≤6) = P (X=1) + P (X=2) + P (X=3) + P(X=4) = P(X=5)

$$= 12.\frac{1}{4}\left(\frac{3}{4}\right)^{11} + \binom{12}{2}\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right)^{10} + \binom{12}{3}\left(\frac{1}{4}\right)^{3}\left(\frac{3}{4}\right)^{9} + \binom{12}{4}\left(\frac{1}{4}\right)^{4}\left(\frac{3}{4}\right)^{8} + \binom{12}{5}\left(\frac{1}{4}\right)^{5}\left(\frac{3}{4}\right)^{7}$$

$$= \frac{3^7}{4^{12}}[972 + 1782 + 1980 + 1485 + 72]$$

$$= \frac{3^7}{4^{12}}(6291) = 0.8201$$

5. Example: it the probability of a defective bolt is 1/8. Find (i) The mean (ii) The variance for the distribution of defective bolt of 640.

Solution: Given the probability of defective bolt is p=1/8 and q=1-p=1-1/8=7/8, n=640

Then mean $\mu = np = 640.1/8 = 80$

Variance σ^2 = n p q =640. 1/8. 7/8 = 70

6. Fit a binominal distribution to the following data

X	0	1	2	3	4	5
F	2	14	20	34	22	8

Solution: Given n=5, n=100

$$\bar{x} = \sum_i \frac{f_i x_i}{N} = \frac{0 + 1*14 + 2*20 + 3*34 + 4*22 + 5*8}{100}$$

$$= \frac{0 + 14 + 40 + 102 + 88 + 40}{100}$$

np = men => 5p = 2.84 => p = 2.84/5 =0.57, q = 1-p =1-0.57=0.43

$$x = 0\ then\ \binom{n}{0}p^0.q^{5-0} = (0.43)^5$$

Expected frequency = N P (x) = 100. $(0.43)^5$ = 1

x =1, 5 (0.57) $(0.43)^4$ *100 =10

x= 2, 10 $(0.57)^2$ $(0.43)^3$ = 26

x = 3, 10 $(0.57)^3$ $(0.43)^2$ = 34

x = 4, 5 $(0.57)^4$ $(0.43)^1$ = 23

x = 5, 5. $(0.57)^5$ = 6

Hence theoretical frequencies are 6, 23, 34, 26, 10, 1

X	0	1	2	3	4	5	
Observed frequencies		2	14	20	34	22	8
Theoretical frequencies	6	23	34	26	10	1	

7. Example: six dace are thrown 243 times. how many times do you expect at least two dice to show a 5 or 6

Solution: Given n=6, n = 243.

The probability p of getting 5 or 6 1/3.i.e., p=1/3, q=1-p-1-1/3=2/3

P (X≥2) = 1-P (X=0) – P (X=1)

Where p(x) is probability of getting x successes among n trial is $\binom{n}{x}p^x.q^{n-x}$

$$P\ (x \geq 2) = 1 - \binom{n}{0}p^0.q^n - \binom{n}{1}p^1q^{n-1}$$

$$=1 - (2/3)^6 – 6.1/3.\ (2/3)^5$$

$$= 1 - \frac{2^6}{3^6}(1+3) = \frac{473}{729}$$

Then expected number of dice = N. P(x) = 243. 473/729=473/3=158.

8. Example: if x is a binomially distributed random variable with E(x) =2 and V=(x) =4/3. Find the distribution of x

Solution: Given the mean and variance of binominal distribution is np=2, npq=4/3

$$q = \frac{npq}{np} = \frac{4/_3}{2}, p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$$

$$with\ p = \frac{1}{3}, q = \frac{2}{3} the\ binomial\ distribution\ is\ \binom{n}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{n-x}$$

Poisson distribution:

Poisson distribution was discovered by the French mathematician and physicist simian denies Poisson in 1837 sometimes we comes across a rare event which occurs once in number of trails for example, consider the event of receiving a telephone calls at a particular telephone exchange in some specified time . if we consider a trial as a number of calls on particular time and the outcome of the trial as to receive a call or not to receive a call then clearly “n” represents the number of calls during a particular time period is very large and it is difficult to find it exactly also the probability “p” of receiving a call is very small however the mean number of calls in the time period is np= λ (say) is finite constant in these

situations is x denotes number of calls then the probability function of random variable in the given time period can be given by

$p(x = x) = \frac{e^{-\lambda^t}\lambda^x}{x!}$Where e is a constant with approximate value 2.7183. This distribution of x is called Poisson distribution and the variable x is called a Poisson vitiate.

Definition:

A random van able x is said to follow a Poisson distribution if it assumes only non-negative integer value and it probability mass function (p.n.f) is given by $p(x = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}; x = 0,1,2, - - - - \lambda > 0 \\ 0, \qquad\qquad otherwise \end{cases}$

There λ is called as the parameter of the distribution Note: if x is a Poisson vitiate with parameter λ . Then it is denoted as $x \sim P(\lambda)$

2. The total probability $\sum_{x=0}^{\infty} P(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}$
$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda}.e^{\lambda} = 1$$

3. distribution function of Poisson distribution is given by
$$F(x) = P(X \geq x) = \sum_{r=0}^{x} P(r)$$
$$= \sum_{r=0}^{x} \frac{e^{-\lambda}.\lambda^2}{r!} = \sum_{r=0}^{x} \frac{e^{-\lambda}.\lambda^2}{r!} \quad x = 0,1,2 - - -$$

4. Poisson distribution is a limiting case of binomial distribution uses or live examples of Poisson distribution :
   The Poisson distribution may be useful in the following instances.
   1. Number of printing mistake as at each page of the book.
   2. Number of telephone calls received at a particular telephone exchange in some unit of time.
   3. Number of subsides reported in a particular city.
   4. Number of air accidents in some unit of time.
   5. Number of defective material in a packing manufacturing company.

   Poisson distribution is a limiting case binomial distribution the Poisson distribution is a limiting case of binomial distribution under the following conditions.

   (i). the number of trials “n” is large i.e., n $\rightarrow \infty$

   (ii). The constant probability of success “p” for each trial is very small, i.e.,p $\rightarrow 0$

   (iii) np is finite , say λ =np

   $$p = \frac{\lambda}{n} \ and \ \ q = 1 - \frac{\lambda}{n}$$

By definition the p.m.f of binomial distribution is

$$P(x) = \binom{n}{x} p^x . q^{n-x}, \qquad x = 0,1,2, - - - - n$$

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

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

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

$$\lim_{n\to\infty} P(x) = \frac{\lambda^x}{x!} \lim_{n\to\infty} \left\{(1 - 1/n)(1 - 2/n) - - - \left(1 - \frac{(x-1)}{n}\right)\right\} \frac{\lim_{n\to\infty} (1 - \lambda/n)^n}{\lim_{n\to\infty} (1 - \lambda/n)^x}$$

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

$$= \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0,1,2, - - - \begin{bmatrix} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda} \\ \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^x = 1 \end{bmatrix}$$

Hence Poisson distribution is a limiting case of binomial distribution.

Mean and variance of Poisson distribution:

The probability mass function of Poisson distribution with parameter λ is given by

$$P(X) = \frac{e^{-\lambda} . \lambda^x}{x!} \quad x = 0,1,2, - - - - \infty$$

$$Mean\ \bar{x} = E(X) = \mu_1^1 = \sum_{x=0}^{\infty} x . P(x) = \sum_{x=0}^{\infty} x . \frac{e^{-\lambda}\lambda^x}{x!}$$

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

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

$$= \lambda .$$

Mean of the Poisson distribution is λ

Variance of the Poisson distribution is

$$V(X) = \sigma^2 = E(X^2) - [E(X)]^2 = \mu_2^1 - (\mu_1^1)^2$$

$$\mu_2^1 = E(X^2) = \sum_{x=0}^{\infty} x^2.P(x) = \sum_{x=0}^{\infty} (x(x-1)+x)\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{(x-2)!} + e^{-\lambda}\sum_{x=0}^{\infty} x.\frac{\lambda^x}{x!}$$

$$= e^{-\lambda}\lambda^2\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} + \sum_{x=0}^{\infty} x.\frac{e^{-\lambda}\lambda^x}{x!}$$

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

$$Var(X) = \sigma^2 = E(X^2) - [E(X)^2] = \lambda.$$

The variance of Poisson distribution is λ .

Hence the .mean and variance of Poisson distribution are equal

Mode of the Poisson distribution:

If x is a non-negative random variable following a Poisson distribution and its p.m.f. is given by

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

Mode is the value of x for which P(x) is maximum

$$\frac{P(x)}{p(x-1)} = \frac{\frac{e^{-\lambda}\lambda^x}{x!}}{\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}} = \frac{\lambda}{x} ----\rightarrow (1)$$

Case (i): when λ is not an integer?

Let us suppose that k is the integral part of λ

$$\frac{P(1)}{P(0)} > 1, ---\frac{P(k-1)}{P(k-2)} > 1, \frac{P(k)}{P(k-1)} > 1,$$

$$and\ \frac{P(k+1)}{P(k)} < 1, \frac{P(k+2)}{P(k+1)} < 1, ------$$

Combining the above expressions in to a single expression, we get P(0),<P(1),<P(2)------ P(K-2)< P(K-1)< P(K)> P(K+1)>P(K+2)-----,

Which shows that P (K) is the maximum value hence in this case the distribution is uni modal and the integral part of λ is the unique modal value.

Case (ii): when integer there we have

$$\frac{P(1)}{P(0)} > 1, \frac{P(2)}{P(1)} > 1, ---- \frac{P(s-1)}{P(s-2)} > 1\ and$$

$$\frac{P(s)}{P(s-1)} = 1, \frac{P(s+1)}{P(s0} < 1, \frac{P(s+2)}{P(s+1)} < 1, - - -$$

P (0) < P (1) < P (2) -------- < P(s-2) < P(s-1) = P(s)>P(S+1)>P(S+2) ---------

In this case we have two maximum values(S-1) and P(s) and thus the distribution is bimodal and two nodes are s-1 and s i.e; at (λ -1) and λ (s= λ)

Note: 1. Sum of the independent Poisson variants is also a Poisson variant. That is, if X_i, (i=1, 2, 3, -----n) are independent Poisson variant with parameters λ_i , i=1,2,3,-----n respectively , then

$\sum_{i=1}^{n} X_i$ is also a Poisson variant with parameters $\sum_{i=1}^{n} \lambda_i$

Symbolically $M_{xi}(t) = e^{\lambda\, i(et-1)}$, i=1,2,3-------n

2. The converse of the above result is also true. i.e.; y X_1, X_2----X_n are independent and

$\sum_{i=1}^{n} X_i$ has a Poisson distribution, then each of the random variables X_1, X_2----X_n has a Poisson distribution. Let X_1 and X_2 be independent random variables so that X_1~P (λ_1) and X_1+X_2 (λ_1+ λ_2) then x_2~P (λ_2) then X_2~P (λ_2).

3. The difference of two independent Poisson variants is not a Poisson variant Which cannot be put in the form hence (X_1-X_2) is not a Poisson variant.

$$i.e, M_{X_1 - X_2} = e^{\lambda_1 (e^t - 1)} + \lambda_{2(e^t - 1)}$$

Which cannot be put in the form $e^{\lambda(e^t-1)}$ hence ($X_1 - X_2$) is not a Poisson variant.

Worked out examples:

Example 1:

2% of the items of a factory are defective the items Are packed in the boxes .What is the probability that there will be (1) 2defective items (2) at least three defective items (3) 2 less than defective items less than 56 in a box of100 items

Solution: Given n=100, p probability of defective items 2/100=0.02 then μ =np= 100 $\times$ 0.02 = 2

(i) X=2

$$P(X, \lambda) = P(2,2) = \frac{e^{-2}(2)^2}{2!} = 2e^{-2} = 2 * 0.136 = 0.272.$$

(ii) P(X≥3) = 1- P(X=0)- P(X=1)- P(X=2)

$$= 1 - e^{-2} - 2e^{-2} - e^{-2}\frac{2^2}{2!} = 1 - 5e^{-2} = 1 - 5 * 0.136 = 0.320$$

(iii) P(2<X<5) = P(X=3)+ P(X=4)

$$\frac{e^{-2}.2^3}{3!} + \frac{e^{-2}.2^4}{4!} = 2e^{-2} = 0.272$$

Example 2:

The probability of Poisson variant taking the values 1 and2 are equal calculate the probabilities of the variant taking the values d 0 and 3.

Solution:

We know that $P(X,\lambda) = \frac{e^{-\lambda}.\lambda^x}{x!}$, $x = 0,1,2,---\infty$

Then from given problem

$$P(1,\lambda) = P(2,\lambda)$$

$$e^{-\lambda}.\lambda = \frac{e^{-\lambda}.\lambda^2}{2!} \qquad \lambda = 2$$

Then P(X=0) = P (0, 2) = e^{-2} = 0.136

$$P(X=3) = P(3,2) = \frac{e^{-2}.2^3}{3!} = \frac{4e^{-2}}{3} = 0.181$$

Example 3:

One fifth percent of the blades produced by a blade manufacturing factory turn out to be defective the blades are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing (i) no defective blade (ii) one defective blade in a consignment of1000.

Solution: Given the Probability of getting no defective blades = 1/500 = p, n= 10

Then $\lambda = np = 10.\frac{1}{500} = \frac{1}{50} = 0.02$

Assuming that X is a Poisson variant with $\lambda = 0.02$

$p(x) = \frac{e^{-\lambda}.\lambda^x}{x!}$

(i) Probability of getting no defective blade is $P(x=0) = e^{-0.02} = 0.9802$

∵ The number of packets containing no defective blades is 10000*0.9802=9802.

(ii) Probability of getting one defective blade $P(x=1) = \frac{e^{-0.02}.0.02}{1!} = 0.0196$.

∵ The number of packets which may containing one defective blade = 0.0196*10000=196

Example 4:

Assuming that the probability of an individual; coalminer being killed in a mine accident during a year is 1/1000. Find the probability that in a mine employing 500 miners there will be at least one accident in a year.

Solution: Probability of an accident of a coal mines =1/100

Then $\lambda = np = \frac{1}{1000}.500 = \frac{1}{2}$.

Probability that there will be at least one accident in year is P (X≥1) = 1- P(X=0) = 1- $e^{-1/2}$.

Example 5:

Given that P(X=2) = 45 P(X=6)-3p(X=4) for a Poisson variant X, find probability that (i) X≥1 (ii) X<2.

Solution: $P(X,\lambda) = \frac{e^{-\lambda}.\lambda^x}{x!}$

Given that P(X=2) = 45, P(X=6)- 3P(X=4)

$$\frac{e^{-\lambda}.\lambda^2}{2} = 45\ \frac{\lambda^6 e^{-\lambda}}{720} - 3.\frac{e^{\lambda}.\lambda^4}{24}$$

$$1 = \frac{\lambda^4}{8} - \frac{\lambda^2}{4} => \lambda^4 - 2\lambda^2 - 8 = 0$$

$$=> (\lambda^2 - 4)(\lambda^2 + 2) = 0$$

$\lambda^2 + 2$can't be possible since the roots are imaginary.

$$\lambda^2 - 4 = 0 => \lambda \pm 2$$

Since mean can't be negative $\lambda = 2$

(i) P(X≥1) = 1- P(X=0) =1-e^{-2} = 1-0.136 = 0.864
(ii) P(X<2) = P(X=0) + P(X=1) = e^{-2} + e^{-2}.2 = e^{-2}.3.
= 3*0.136 = 0.408.

Example 6: If two cards are drawn from a pack of 52 cards which are diamonds using Poisson distribution final the probability of getting two diamonds at least three times in 51 consecutive trials of two cards drawing each time

Solution: Let P=Probability of getting two diamonds from a pack of 52, if two cards are selected randomly = $13c_2/52c_2$ = 3/51 n=51

$$Then\ \lambda = np = \frac{3}{51}.51 = 3$$

Probability of getting two diamonds at least three times is

P (X≥3) = 1- P(X=0) - P(X=1) - P(X=2)

$$= 1 - e^{-3} - 3.e^{-3} - e^{-3}.\frac{9}{2}$$

$$= 1 - e^{-3}\left(\frac{17}{2}\right) = 0.5762$$

Note: We know that probability at any value of x when mean is given by λ is

$P(X,\lambda) = \frac{e^{-\lambda}.\lambda^x}{x!}$ Pmf of Poisson distribution

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

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

∵ The probability at x = x = 1 is λ / (x=1) times the probability at x

Example 7: If x is a Poisson variant such that P(X=0) find P(X=0) and using recurrence formula find probability at x=1, 2, 3, 4 and 5

Solution: given probability at x=0= probability at x=1

i.e., P(x=0) = P(x+1)

Then we have from definition of Poisson distribution

$$P(X,\lambda) = \frac{e^{-\lambda}.\lambda^{x}}{x!}$$

$=> e^{-\lambda} = \lambda . e^{-\mu}$ (at P (X=0) & P (X=1))

$\lambda = 1$

$$P(X=0) = e^{-\lambda} = e^{-1} = 0.368 = P(X=1)$$

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

Put X=1

P(X=2) = ½ P (1) = ½ 0.368 = 0.184

Put x=2

P(X=3) = 1/3 P (2) = 0.0613

P (4) = ¼ P (3) = 0.0153

P (5) = 1/5 P (4) = 0.00306

∵ Probabilities at x=0, 1, 2, 3, 4, 5 are 0.368, 0.368, 0.184, 0.0613, 0.0153, 0.00306.

Example 8: it a Poisson distribution for the following data and calculated frequencies

X	0	1	2	3	4
F(x)	109	65	22	3	1

Solution: Here N=200

Mean of the Poisson distribution = λ

$$To\ find\ \lambda, \qquad \lambda = \frac{\sum_i f_i x_i}{\sum_i x_i} = 0.61$$

Frequency f(x) N P(x) = 200 P(x)

P(0) =$e^{-\lambda} = e^{-0.61} = 0.5435$

f (0) = 200 * 0.5435 = 109

f (1) = 200 * P (1)

$$f(x+1) = N.P9x+1) = \frac{N.\lambda}{x+1}P(x) = \frac{\lambda}{x+1}f(x)$$

f (1) = 0.61. f (0) = 109 *0.61 = 65

f (2) = 0.61/2 f (1) = 0.305 * 65 = 20

f (3) = 0.61/3 f (2) = 20 *0.61/3 = 4

f (4) = 0.61/4 f (3) = 0.61/4 *4 = 0.61 ≈ 1

Expected frequencies are

x	0	1	2	3	4
f(x)	109	65	22	3	1

Simulation of Discrete Distribution:

Simulation is a method of solving decision making problems by designing constructing and manipulation a mode of the real system .it is defined to be the action of performing experiments on a model of a given system it duplicates the essence of a system or activity without actually obtaining the reality.

The Monte Carlo Technique has become so much a part of simulation models that the terms are other assumed to be synonymous it is however only a technique writhen simulation the procedure of Monte Carlo involves the selection of random observations writhen the simulation model and consists of the following two steps.

Step 1: Generate random observations from a uniform distribution on (0, 1)

Step 2: Using step 1, generate random observations from any desired probability distribution.

Random Observations Generation:

To generate a sequence of random observations from any probability distribution, when a sequence of random numbers is known as the fundamental and most simple technique is the so called inverse probability transformation method or generating by inversion the procedure involves the following three steps.

Step 1: Construct the cumulative distribution function $F_X(X)$ OF The Random Variable X.

Step 2: Generate normalized (0, 1) random numbers U_1,U_2-------

Step 3: Set $F_X(X)$ equal to the random decimal number and solve for x. the value of x these obtained will be the desired random observation from the probability distribution.

To obtain x_1, the first random observation corresponding to $F_X(X)$we similarly enter the intonate with U_1,project over and down as that shown in fig 1 .then the resulting abscess value will be $x_1$repitition of the procedure with U_2,U_3----- WILL GIELD X_2,X_3-------

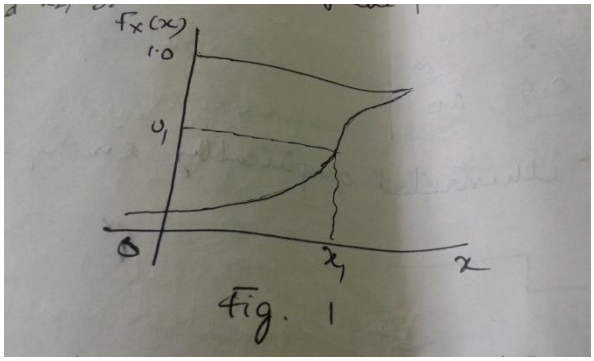


Fig. 1

Now, $F_X(X)$ is uniform over the interval (0, 1) regardless of the distribution of x. If $F_X(X)$ is continuous and strictly increasing then when a number u on [0, 1], there is unique value for x such that F(x) =U .Symbolically this value of x is denoted by $F^{-1}_X(U)$.The problem now is to generate a sequences of uniform random numbers U_n, n=1, 2------and from these determine the associated observations

$$F_X(x_n) = U_n \Rightarrow x_n = F_X^{-1}(U_n)$$

For some probability distribution such as uniform and exponential it is not necessary to plot $F_X(X)$ to obtain x_c, since $F^{-1}(U_i)$ can be obtained analytically. For other distributions, such as normal and binomial analytical inversion not possible but other methods can be used to obtain $F_X^{-1}(U)$

Simulating a discrete distribution and for those probability distributions which cannot be presented in closed form tabular method is generally used for example consider following probability density function .

x	0	1	2	3
P(x)	0.4	0.3	0.2	0.1

The C.d.f is given by

X	0	1	2	3
F(x)	0.4	0.7	0.9	1.0

The above data is illustrated graphically in fig 2

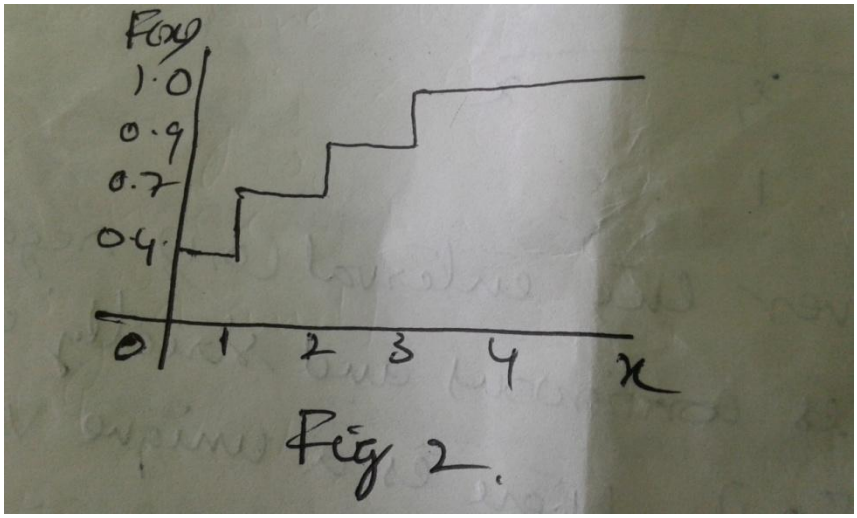


Fig 2.

Random numbers can be categorized to define the random viriate x_n uniquely as follows

$$x_n = \begin{cases} 0 \le u \le 0.4, & for\ n = 0 \\ 0.4 < u \le 0.7, & for\ n = 1 \\ 0.7 < u \le 0.9, & for\ n = 2 \\ 0.9 < u \le 1.0, & for\ n = 3 \end{cases}$$

This implies that for any discrete random variable X the random variant assumes the value I is the random number is such that

$F(j-1) \le U \le F(j)$

Bionomical distribution:

For example if we consider binomial distribution with parameters n and p then

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0,1, - - - - n$$

X is the number of successes in n independent Bemoulli trials each with success probability P. then generate n Bemoulli (P) random variables y_1------y_n

Set $x = Y_1 + Y_2 + \text{-----} + Y_n$

This can also be derived by using the following results let Y_1 , Y_2---- be independent geometries (P) random variables and I the smallest index such that

$$\sum_{i=1}^{I+1} (y_i + 1) > n$$

Then the index I has a binomial distribution with parameter n and P.

And also let Y_1 , Y_2-----be independent exponential random variable with mean 1 the smallest index such that

$$\sum_{i=1}^{I+1} \frac{Y_i}{n-i+1} > -L_n(1-p)$$

Then the index 1 has a binomial distribution with parameters n and p.

Poisson distribution:

Let = $x_1 + x_2 + \text{-----} + x_n$ be identical exponential variables with mean $\frac{1}{\lambda}$. Define n such that

$x_1 + x_2 + \text{----} x_n \leq t\ x_1 + x_2 + \text{----} + x_{n+1}$ ----------→(1)

Then the distribution of n will be Poisson with λ t the relationship then this can be used to generate random observation from Poisson distribution.

Now using the exponential formula

$$x_i = \frac{1}{\lambda} \log \frac{1}{u_i}$$

When U_i is a normalized (0, 1) random number substituting it in the along inequality we get

$$\sum_{i=1}^{n} \frac{1}{\lambda} \log \frac{1}{u_i} \leq t < \sum_{i=1}^{n+1} \frac{1}{\lambda} \log \frac{1}{u_i}$$

Or

$$-\sum_{i=1}^{n} log u_i \leq \lambda t < -\sum_{i=1}^{n+1} log u_i$$

Or

$$\log \prod_{i=1}^{n} u_i \geq -\lambda t > \log \prod_{i=1}^{n+1} u_i$$

Taking exponentials of equation which yields

$$\prod_{i=1}^{n} u_i \geq e^{-\lambda t} > \prod_{i=1}^{n+1} u_i$$

To illustrate the use of the formula suppose the poisson distribution has mean λ t=3 then e $-\lambda t$ =0.04979.

The problem then is to generate random numbers unitil the inequalities are satisfied consider U_1=0.09656, U_2=0.96657, U_3=0.64842 and U_4=0.49922.since U_1 U_2 U_3 =0.06051 and U_1 U_2 U_3 U_4=0.03021, then the inequalities are satisfied for n=3 which is a random vacate having the given Poisson distribution

Continuous distribution:

Exponential distribution:

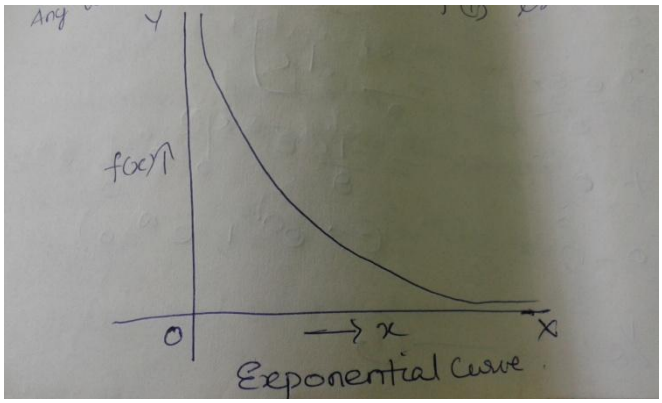
In the theory of continuous distributions we commonly say normal distribution is an example of model for any data because its range is an spread over ($-\infty$ $+\infty$),for data that is generally positive valued application of normal distribution is not suitable sometimes then the simple model that can be useful is exponential distribution the random variable of exponential distribution is positive valued .like normal distribution it has many smooth properties it can be used on a good model for life time of number of industrial products. in this section we study a theoretical and practical expects of exponential distributions.

Definition:

A continuous random variable 'x'is said to follow an exponential distribution with parameter is its probability density function is given by

$$f(x) = \begin{cases} \theta\, e^{-\theta x}, & 0 < x < \infty, \theta > 0 \\ 0, & Otherwise \end{cases} \qquad \text{--------}\rightarrow(1)$$

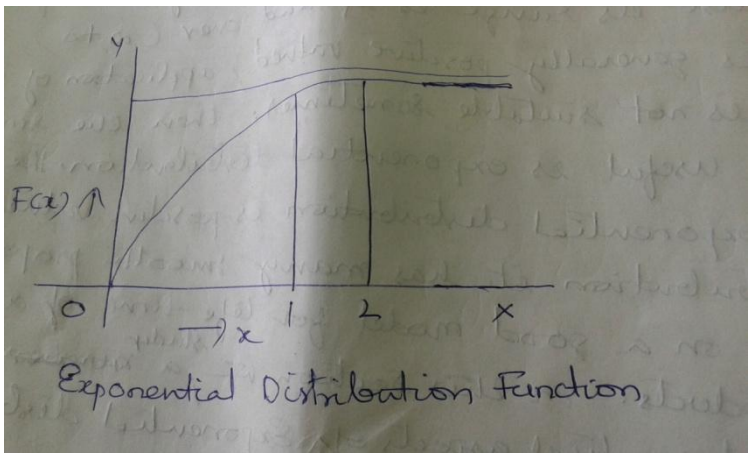
Any variate having the p.d.f (1) is expressed as x~Expo (θ)



The cumulative distribution function F(x) is given by

$$F(x) = \int_0^x \theta e^{-\theta t}\, dt = 1 - e^{-\theta x}, \qquad x > 0$$

$$F(x) = \begin{cases} 1 - e^{-\theta x}, & x > 0 \\ 0, & Otherwise \end{cases}$$



Mean and variance of exponential distribution:

We have Mean = $\mu_1 = E(x) = \int_0^\infty x.f(x)dx$

$= \int_0^\infty x.\theta.e^{-\theta x}\,dx\,\left(f(x) = \theta.e^{-\theta x}\right)$

$$= \theta \int_0^\infty x.e^{-\theta x}\,dx$$

$$= \theta\left[x.\int_0^\infty e^{-\theta x}\,dx - \int_0^\infty \left[\frac{d}{dx}x\right].\int e^{-\theta x}\,dx\right]$$

$$= \theta.\left[\frac{x.e^{-\theta x}}{-\theta}\int_0^\infty \quad + \int_0^\infty \frac{e^{-\theta x}}{\theta}dx\right]$$

$$= 0 + \int_0^\infty e^{-\theta x} = \frac{-1}{\theta}.e^{-\theta x}\int_0^\infty \quad = \frac{-1}{\theta}(e^\infty - e^0)$$

$$= \frac{-1}{\theta}(0-1) = \frac{1}{\theta}$$

$$\mu_1 = E(x) = \frac{1}{\theta}.$$

Variance V(X) = $E(X^2) - [E(X)]^2$ or $\mu_2 = \mu_2{}^1 - (\mu_1{}^1)^2$

But E(X) = 1/θ and $E(X^2)$ is given by

$$E(X^2) = \mu_2^1 = \int_0^\infty x^2.f(x)dx = \int_0^\infty x^2.\theta.e^{-\theta x}\,dx$$

$$= \theta.\int_0^\infty x^2.e^{-\theta x}\,dx = \theta\left[x^2\int_0^\infty e^{-\theta x}\,dx - \int_0^\infty \left[\left(\frac{d}{dx}x^2\right)\int_0^\infty e^{-\theta}\,dx\right]dx\right]$$

$$= \left[x^2.e^{-\theta x}\int_0^\infty \quad + \int_0^\infty 2x.e^{-\theta x}\,dx\right]$$

$$= 0 + 2\left[x.\int_0^\infty e^{-\theta x}\,dx - \int_0^\infty \left[\left(\frac{d}{dx}x\right).\int e^{-\theta x}\,dx\right]dx\right]$$

$$= 2.\frac{x.e^{-\theta x}}{-\theta}\int_0^\infty + 2.\int_0^\infty \frac{e^{-\theta x}}{\theta}dx$$

$$= 0 + 2.\frac{e^{-\theta x}}{-\theta^2}\int_0^{\infty}$$

$$E(x^2) = \frac{-2}{\theta^2}(e^{-\infty} - e^0) = \frac{-2}{\theta^2}(-1) = \frac{2}{\theta^2}(\because e^{-\infty} = 0, e^0 = 1)$$

$$\therefore \mu_2^1 = E(x^2) = \frac{2}{\theta^2} \quad ---\rightarrow (2)$$

Hence $V(x) = E(x^2) - [E(x)]^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$

$$\therefore \mu_2 = V(x) = \frac{1}{\theta^2}$$

Moment generating function:

The moment generating function (m.d.f) of a random variable x (about again) having the probability function f(x) is given by

$$M_X(t) = E(e^{tx}) = \int e^{tx}.f(x)dx$$

$\because$ x~expo ( )we have

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

$$= \theta.\frac{e^{-(\theta - i)x}}{-(\theta - t)}\int_0^{\infty}$$

$$= \frac{\theta}{-(\theta - t)}(e^{-\infty} - e^0) = \frac{\theta}{(\theta - t)} = \left(1 - {}^{t}/_{\theta}\right)^{-1}$$

Using the binomial expansion we get

$$M_X(t) = \sum_{r=0}^{\infty}\left(\frac{t}{\theta}\right)^r$$

$$\therefore \mu_r^1 = E(X^2) = Coefficient\ of\ \frac{t^r}{r!}\ in\ M_X(t)$$

$$= \frac{r!}{\theta^r};\ r = 1,2, - -$$

Note: 1.Pearson`s measure of skewers is $\gamma_1 = 2 = \sqrt{\beta_1}$

2. Pearson`s measure of kurtosis is $\gamma_2 = \beta_2 - 3 = 6$

3. Medan of exponential distribution is given by $m = \theta^{-1}\log_e 2$.

4. The relation between exponential and uniform distributer is, if $X \sim exp(\theta),\ then\ Y = e^{-\theta x}\ is\ U(0,1)$.

5. Memory less property of exponential distribution is given by

$$P(Y \leq x \,/\, X \geq a) = \mathrm{P}(\mathrm{X} \leq \mathrm{x})$$

Worked out Examples:

Example 1: Show that for the exponential distribution

$$\mathrm{dP(x)} = y_0 . e^{-x/\sigma} dx, 0 \leq x \leq \infty;\ \sigma > 0, (y_0 is\ a\ constant)$$

mean and S.D are equal.

Solution: In order to change the given distribution in to probability density function , we must have

$$\int_0^\infty y_0 . e^{-x/\sigma} dx = 1\ (\therefore the\ total\ area\ under\ curve\ is\ unity)$$

$$\Rightarrow y_0 \int_0^\infty e^{-x/\sigma} dx = 1$$

$$\Rightarrow y_0 . \sigma . \left[-e^{-x/\sigma}\right]_0^\infty = 1 \Rightarrow y_0 . \sigma[-(e^{-\infty} - e^0)] = 1$$

$$\Rightarrow y_0 . \sigma = 1 \Rightarrow y_0 = 1/\sigma$$

Hence the P d f $\mathrm{f(x)} = \frac{1}{\sigma} e^{-x/\sigma};\quad 0 \leq x \leq \infty, \sigma > 0$

Mean $= \mathrm{E(x)} = \mu_1^1 = \int_0^\infty x . f(x) dx = \int_0^\infty x . \frac{1}{\sigma} e^{-x/\sigma}\, dx$

$$= \left[x . \frac{e^{-x/\sigma}}{-1/\sigma} \right]_0^\infty - \frac{1}{\sigma} \int_0^\infty \frac{e^{-x/\sigma}}{\frac{-1}{\sigma}} dx (by\ integrating\ by\ parts)$$

$$= 0 + \int_0^\infty e^{-x/\sigma} dx$$

$$= \left(\frac{e^{-x/\sigma}}{-1/\sigma} \right)_0^\infty = -\sigma(e^{-\infty} - e^0) = \sigma(\because e^{-\infty} = 0, e^0 = 1)$$

$$\mu_1^1 = \sigma ------(1)$$

Similarly

$$\mu_2^1 = \int_0^\infty x^2 . f(x) dx = \int_0^\infty x^2 . \frac{1}{\sigma} e^{-x/\sigma} dx$$

$$= \frac{1}{\sigma} \int_0^\infty x^2 e^{-x/\sigma} dx$$

$$= \frac{1}{\sigma}\left[x^2 . \frac{e^{-x/\sigma}}{-1/\sigma} \right]_0^\infty - \frac{1}{\sigma}\int_0^\infty \left[2x . \frac{e^{-x/\sigma}}{-1/\sigma} \right]_0^\infty dx$$

$$= 0 + 2.\int_0^\infty e^{-x/\sigma} dx$$

$$= 2.x\left| \frac{e^{-x/\sigma}}{-1/\sigma} \right|_0^\infty - 2\int_0^\infty \frac{e^{-x/\sigma}}{-1/\sigma} dx$$

$$= 0 + 25\int_0^\infty e^{-x/\sigma} dx$$

$$\mu_2^1 = 2\sigma \frac{e^{-x/\sigma}}{-1/\sigma}|_0^\infty = 2\sigma^2\left[-(e^{-\infty} - e^0) \right] = 2\sigma^2 \text{ -----(2)}$$

$$\therefore V(X) = \mu_2 = \mu_2^1 - \mu_1^{2 1} = E(x^2) - \left[E(x)\right]^2 = 2\sigma^2 - \sigma^2 = \sigma^2 (\because \textit{from}(1), \textit{and}(2))$$

Hence standard deviation (S.D)= $\sqrt{v(x)} = \sqrt{\sigma^2} = \sigma$

Therefore the mean and s.d are equal to σ.

Example 2: The water consumption of a city in excess of 20,000 gallows, is exponentially distribution with mean 20,000.the city`s water works has a daily stock of 40,000 gallons .what is the probability that the stock is insufficient for at least two of the three days selected at random ?

Solution : if y is the total consumption in a day then x=y-20,000 has an exponential distribution with mean 20,000 with the

$$f(x) = \frac{1}{20{,}000} e^{-x/20{,}000}, \textit{for} 0 \le x \le \infty$$

Since if the demands exceeds 40,000 gallons then the stock

Will be proved insufficient

$$x \ge 40{,}000 - 20{,}000$$

i.e., $\Rightarrow x \ge 20{,}000$

The probability that the stock remains insufficient on any particular day is given by

$$P(x \ge 20{,}000) = \int_{20{,}000}^\infty \frac{1}{20{,}000} e^{-x/20{,}000} dx = e^{-10}$$

The probability that the stock is insufficient for at least two of three days selected at random is equal to sum of probability that it is insufficient for all the three days and probability that it is inefficient for two of the three days.

$$= (e^{-10})^3 + 3_{c_2}(e^{-10})^2(1-e^{-10})$$
$$= e^{-10} + 3.e^{-20}(1-e^{-10})$$
$$= e^{-20}(e^{10} + 3(1-e^{-10})) = e^{-20}(e^{-10} + 3 - 3e^{-10})$$
$$= e^{-20}(3-2e^{-10})$$

Normal Distribution

The Normal Distribution was introduced in 1733 by Mathematician De-moivre,who obtained this continuous distribution as a limiting case of the binomial distribution and apptead it to problems arising in the game of chance. Later Laplace and gauss derived it independently of each other as the distribution of errors in physical measurements. These the normal distribution has got wide applications in the theory of statistics

Definition :

A random variable X is said to have a normal distribution with parameters μ,called mean and the variance if its density function is given by the probability law

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}.\exp\left[\frac{-1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right]$$

Or

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

Therefore, random variable X with mean μ and variance σ^2 following the normal distribution is expressed as $x \sim N(\mu,\sigma^2)$ if $x \sim N(\mu,\sigma^2)$, then $z = \frac{X-\mu}{\sigma}$ is a standard normal variate with mean o i.e., E(Z)=0 and variance 1,i.e., v(z)=1 and is denoted by z~n(0,1)

Hence the probability density function of standard normal variete Z is given by

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

And the corresponding distribution function , denoted by $\phi(z)$ is given by

$$\phi(z) = P(Z \le z) = \int_{-\infty}^{z}\phi(u)dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-u^2/2}du.$$

Note : 1. $\phi(-z) = 1-\phi(z)$

2. $P(a \le x \le b) = \phi\left(\frac{b-\mu}{\sigma}\right) where, X \sim N(\mu,\sigma^2)$

Mean and variances of normal distribution

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

Put $\frac{x-\mu}{\sigma}-z\Rightarrow x=\mu+\sigma Z, dx=\sigma dz$

If $x=\infty, Z=\infty, x=-\infty, z=-\infty$

$$E(Z)=\int_{-\infty}^{\infty}(\mu+\sigma z)\frac{1}{\sigma\sqrt{2\pi}}e^{-z^2/2}\sigma.dz$$

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

$$E(Z)=\frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-z^2/2}.dz+\int_{-\infty}^{\infty}Z.e^{-z^2/2}dz$$

$=\mu.1+0$

$=\mu.$

$\because \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-z^2/2}.dz=1$, and the second integral.

$\because \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}ze^{-z^2/2}.dz=0$ is an odd function

To find the variances σ^2 of the normal distribution

$$V(X)=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu)^2 e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma^2 z^2 e^{-z^2/2}\sigma dz(\because \frac{x-\mu}{\sigma}=z \;\&\; dx=\sigma dz$$

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

$$=\frac{2\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z^2 e^{-z^2/2}dz(\because z^2.e^{-z^2/2} is even)$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{2t} e^{-t} dt \left(\frac{z^2}{2} = t\right)$$

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

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t} t^{\frac{3}{2}-1} dt$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \gamma\left(3/2\right) = \frac{2\sigma^2}{\sqrt{2\pi}} \frac{1}{2} \gamma\left(1/2\right) (\because \gamma(n+1) = n\gamma(n))$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{\pi} (\because \gamma\left(1/2 = \sqrt{\pi}\right)$$

$V(X) = \sigma^2$

S.D = $\sqrt{V(x)} = \sqrt{\sigma^2} = \sigma$

Mode of normal distribution :

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

$\therefore f^1(x) = 0 and, f^{11}(x) = -ve$ at that value of x

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

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

=0, when x= μ, since $e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \neq 0$

$$f^{11}(x) = \frac{1}{\sigma^2\sqrt{2\pi}}\left[\frac{-1}{6}\left(\frac{x-\mu}{\sigma}\right)^2 e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} - \frac{e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma}\right] = -ive$$

When x = μ

Hence μ is the mode of normal distribution.

Median of normal distribution :

If M is the median of the normal distribution we have

$$\int_{-\infty}^{m} f(x)dx = \frac{1}{2} \Rightarrow \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{M}\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{M}\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx + \frac{1}{\sigma\sqrt{2\pi}}\int_{M}^{M}\exp(-\frac{(x-\mu)^2}{2\sigma^2}dx = \frac{1}{2} ------(1)$$

But

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{M}\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx + \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{0}\exp(\frac{-z^2}{2})dz = \frac{1}{2}$$

$\therefore$

From (1) we get

$$\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}}\int_{M}^{M}\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx = \frac{1}{2}$$

$\therefore$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}}\int_{M}^{M}\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx = 0 \Rightarrow \mu = M$$

$\therefore$

Hence median of the normal distribution is M= μ

Moment generating Function of normal distribution :

The m.g.f. (about origin)is given by

$$M_x(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]dx$$

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

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

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

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

$$= e^{\mu t + \frac{t^2\sigma^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{-u^2}{2}\right)du$$

Hence $M_x(t) = e^{\mu t + \frac{t^2\sigma^2}{2}}$

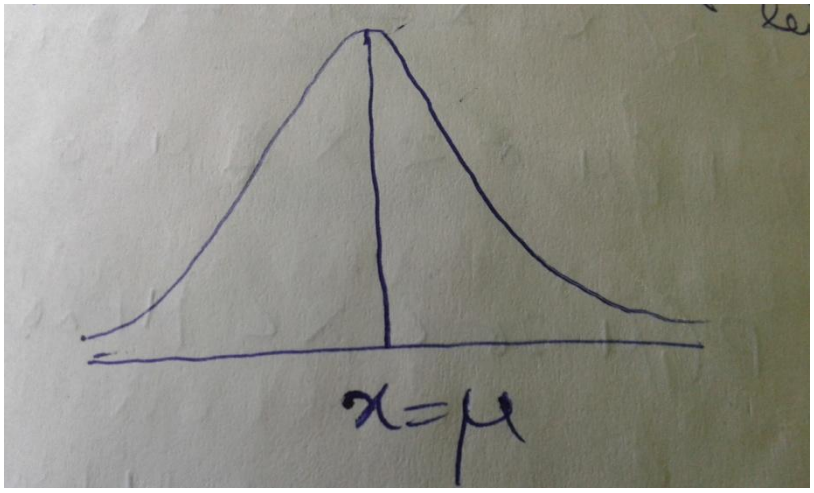
Properties of normal distribution :

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

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

And has the following properties.

1. The curve is bell shaped and synuemetric about the line x= μ



2. Mean median and mode of the distribution coincide at x= μ
3. Q.D.:M.D:S.D:: $\frac{2}{3}\sigma : \frac{4}{5}\sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1 :: 10 : 12 : 15$
4. As x evieases numerically f(x)decreases rapidly the maximum probability occurring at the point x= μ and given by

   $\left[p(x)\right]_{\max} = \frac{1}{\sigma\sqrt{2\pi}}$
5. Area under the normal curve is unity

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

6. Since f(x) being the probability which will never be negative their fore no portion of the curve line below the x-axis
7. X-anis is asymptote to the curve
8. The distribution has points of inflexion at x= μ + and. $f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}$ that is they are equidistant (at a distance) from the mean.
9. Mean deviation from the mean for normal distribution i.e., M.D.(about mean)=4/5 σ .(approximately)
10. Area property of normal distribution
    (i) $P(\mu - \sigma < X < \mu + \sigma) = P(-1 < z < 1) = 0.6826$
    (ii) $P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < z < 2) = 0.9544$
    (iii) $P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < z < 3) = 0.9973$
11. Linear combination of independent normal variates is also a normal variate .

Importance of normal distribution :

This was an importants distribution which was initially discovered for studying the random errors of measurements ,that is during the calculators of orbits of celestial bodies. It happened because of a remarkable coincidence that normal distribution has a wide range of

applications these days in the theory of stastics .To country a few are industria l quality control ,test of the significance sampling distribution of various stastics , graduation of normal curves etc. length of the leaves from a particular point of time, weights of trees of the same variety, weights taken from the group of students, taken of the same age,intelligence,proportion of male to female births for some particular geographical region over a period of years and many other examples from various fields can be given which are studied through normal distribution. Some facts of normal distribution . Some facts opf the normal distribution detailed below.

1. Normal distribution approximates the p d f is of the most of the commonly occurring distribution such as binomial ,Poisson ,Hyper geometric……etc;
2. Many of the sampling distributions such as student t , seducer's F ,parson `s x^2 etc----,are asymptotically normal .Also most of the sampling distributions tend to normality as
3. Sometimes a non normal variat begins to exhibit normality properties under suitable trance formations
4. $P\{|Z| \geq 1.96\} = 0.05, and, P\{|Z| \geq 3\} = 0.0027, if Z is N(g1)$
   These properties of N (0,1) are from the basis of "Large Sample Theory".

5. For a sample size >30 can always be treated as normal , even though parent population is non normal (central limit theorem)
6. In. tests of significance the population is assumed to be normal
7. Normal distribution finds large applications in satisfied quality control in industries and graduation of non normal curves

Worked out Examples :

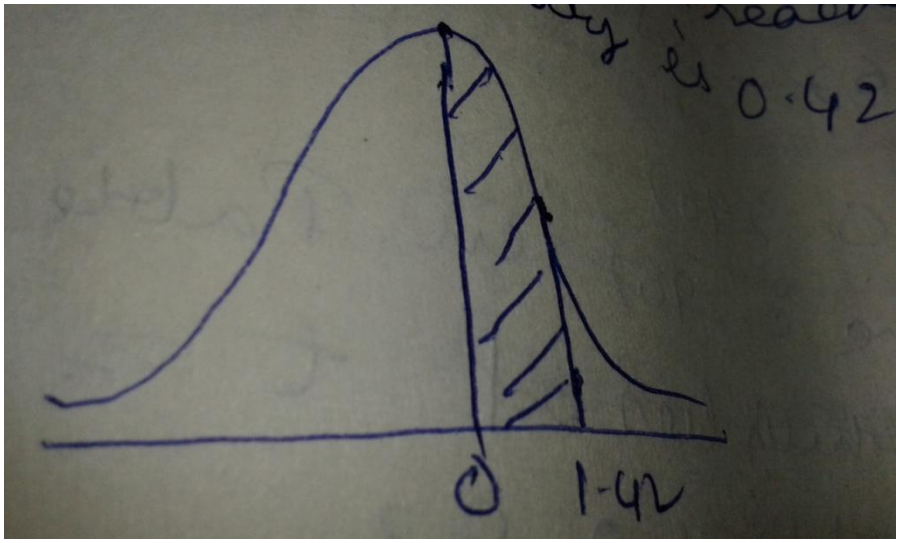
Example : 1.

Let X be a random variable with the standard normal . Find (i)

$P(0 \leq X \leq 1.42)$ (ii) $P(-0.73 \leq X \leq 0)$

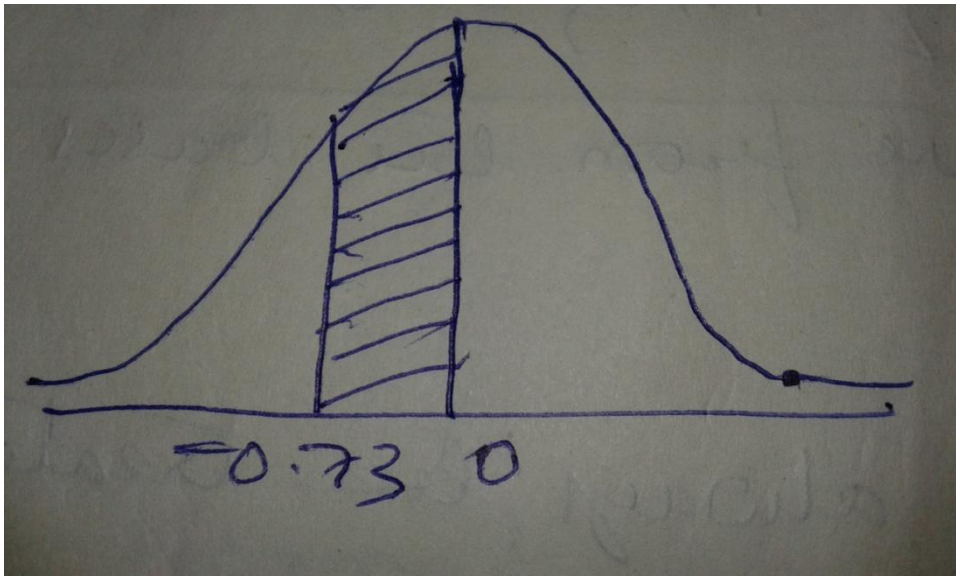
(iii) $P(X \geq 1.\beta$ (iv).Determine the value of t if $P(X \leq t) = 0.7967$

Solution : (i).Here $P(0 \leq X \leq 1.42$ is equal to the aria under the standard normal curve between 0 and 1.42. T hus in table of areas under standard normal curve , look down the first column until 1.4 reached and then continue right to column 2. The entry is 0.4222 .



Hence $P(0 \le X \le 1.42$

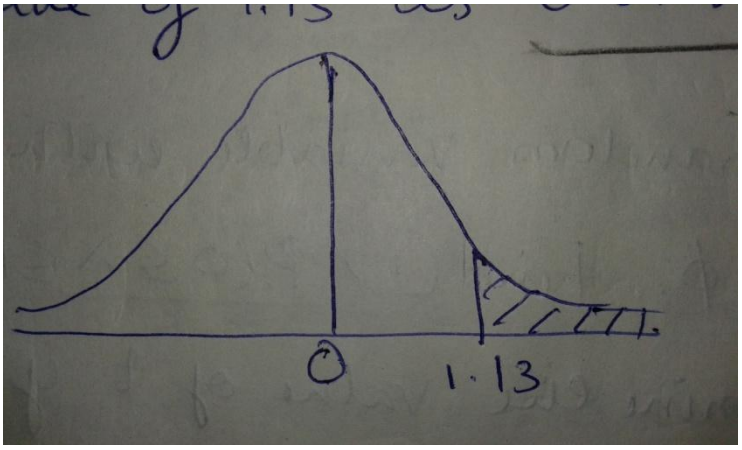
(ii) $P(-0.73 \le X \le 0) = p(0 \le X \le 0.73$ (by symmetry) to the area under the standard normal curve between 0 and 0.73. Thus in table of areas under standard normal curve lookdown the first column until 0.73 reached and then continue right to column 2.the entry is 0.2673.Hence $P(0 \le X \le 0.73) = 0$



(iii) $P(X \ge 1.13$ which can be return has

$$P(X \ge 1.13) = P(X \ge 0) - P(0 \le X \le 1.13$$
$$= 0.5000 - 0.3708 = 0.1292$$

From the table of area under the standard normal curve gives the value of 1.13 as 0.3708 & $P(X \ge 0) = 0.5000$



(iv) $P(X \le t) = 0.7967$

Here t must be positive since the probability is greater than 1/2.therefore we write

$$P(0 \le X \le t) = P(X \le t) - \frac{1}{2}$$
$$= 0.7967 - 0.5000 (\because P(X \le t) = 0.7967$$
$$P(0 \le X \le t) = 0.2967$$

Now observing the value 0.2965 in table of areas under standard normal curve which lies at t = 0.83

We obtain the value of t as 0.83

Example :2. If mean μ =70 and standard deviation is 16 find (i) $P(38 \le X \le 46)$ (ii) $(82 \le X \le 94)$ (iii) $(62 \le X \le 86)$

Solution : given mean μ =70, S.D $\sigma = 16$ Then $Z = \frac{x - \mu}{\sigma}$

(i)
$$x_1 = 38, Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{38 - 70}{16} - 2$$
$$x_2 = 46, Z_2 = \frac{x_2 - \mu}{\sigma} = \frac{46 - 70}{16} = \frac{-24}{16} = -1.5$$

Then
$$P(38 \le X \le 46) = P(-2 \le Z \le -1.5)$$
$$= P(-2 \le Z \le 0) - P(-1.5 \le Z \le 0)$$
$$= P(0 \le Z \le 2) - P(0 \le Z \le 1.5) by$$
By symmetry
=0.4772-0.4332 (from areas of normal tables)
=0.0440 $\because \phi(2) = 04772, \phi(1.5) = 0.4332$

(ii)

$$x_1 = 82, Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{82-70}{16} = \frac{12}{16} = 0.75$$

$$x_2 = 94, Z_2 = \frac{x_2 - \mu}{\sigma} = \frac{94-70}{16} = \frac{24}{16} = 1.5$$

$$P(82 \le X \le 94) = P(Z_1 \le Z \le Z_2)$$

$$= P(0.75 \le Z \le 1.5)$$

$$= P(0 \le Z \le 1.5) - P(0 \le Z \le 0.75)$$

$$= 0.4332 - 0.2734 = 0.1598$$

$(\because \phi(1.5) = 0.4332, \phi(0.75) = 0.2734)$

From areas of normal tables.

(iii) $P(62 \le X \le 86)$

$$x_1 = 62, Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{62-70}{16} = -0.5$$

$$x_2 = 86, Z_2 = \frac{x_2 - \mu}{\sigma} = \frac{86-70}{16} = 1.0$$

$$P(-0.5 \le Z \le 1.0) = P(-0.5 \le Z \le 0) + P(0 \le Z \le 1.0)$$

$$= P(0 \le Z \le 0.5) + P(0 \le Z \le 1.0)$$ (by symmetry)

=0.1915+0.3413=0.5328

$(\because \phi(0.5) = 0.1915, \phi(0.5) = 0.343$

From areas of normal tables.

Example 3. Student of a class were given an examination this marks were found to be normally distributed with mean 55 marks and standard deviation 5.
Find the number of student who is the marks more than 60 is 500 students were written the examination

Solution : Mean μ =55 , S.D. σ =5

$$Z = \frac{x - \mu}{\sigma}, x = 60, then, Z = \frac{60-55}{5} = 1$$

$$P(X > 60) = P(Z > 1)$$

$$= 0.5 - P(0 \le Z \le 1)$$

$$= 0.5 - 0.3413 = 0.1587$$

$(\because \phi(1) = 0.3413,$

From areas of normal tables.

The numbers of students who get more than 60% OF MARKS =0.1587$\times$500=79

Hence 79 students get more than 60% of marks

Example : 4. In a test on 2000 electric bulbs, it was found that the life of a particular make, was normally distributed with an average life of 2040 hours and S.D of 40 hrs estimate the number of bulbs likely to burn for

(i) More than 2140 hrs

(ii) Between 1920 and 2080

(iii) Less than 1960 hrs

Solution : Mean μ =2040hrs, S.D. σ =40, $Z = \frac{x-\mu}{\sigma}$ (i).

(i) X = 2140

$$Z = \frac{2140-2040}{40} = \frac{100}{40} = 2.5$$

$$P(X > 2140) = P(Z > 2.5)$$

$$P(Z > 2.5) = 0.5 - P(0 \le Z \le 2.5)$$

=0.5-0.4938 $(\because \phi(2.5) = 0.4938)$

=0.0062

The number of bulbs likely to burn more than 2140hrs = 0.006 $\times$200=12 bulbs

(ii)

$$x_1 = 1920, Z_1 = \frac{x_1-\mu}{\sigma} = \frac{1920-2040}{40} = \frac{-120}{40} = -3$$

$$x_2 = 2080, Z_2 = \frac{x_2-\mu}{\sigma} = \frac{2080-2040}{40} = 1.0$$

$$\therefore P(1920 \le x \le 2080) = P(-3 \le Z \le 1)$$

$$= P(-3 \le Z \le 0) + P(0 \le Z \le 1)$$ by symmetry

$$= P(0 \le Z \le 3) + P(0 \le Z \le 1)$$

$$= 0.4986 + 0.3413 = 0.8399$$

From normal tables

Number of bulbs likely to burn between 1920 hrs and 2080=0.8399X2000=1679.8, balbs=1680

Number of bulbs likely to burn less than 1960 hrs and 2080=1680.

(iii) $P(X \le 1960)$

X = 1960, $Z = \frac{1960-2040}{40} = \frac{-80}{40} = -2$

$$P(X \le 1960) = P(Z \le 2)$$

$$= 0.5 + P(0 \le Z \le 2)$$

$$= 0.5 + 0.4772$$

$$= 0.9772$$

From normal tables.

Number of bulls likely to turn less than 1960 hrs =0.9772X2000=1954.

Waybill distribution

The distribution is named after waloddi weibull,a Swedish physicist who used it in 1939 to represents the distribution of the of the breaking strength of materials kao ,J.H.K.(1958-59) advocated the use of this distribution in reliability studies and quality control work it is also used a a tolerance distribution in the analysis of question response data

Definition ;

If `X` is a continuous random variable with parameters α and β and follows an weibull distribution according to probability lw has the p.d.f. is given by

$$f(x)=\begin{Bmatrix}\alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, x>0,\alpha>0,\beta>0\\ 0, otherwise\end{Bmatrix}$$

the probability that a random variable having the weifull distribution will take on a value less than , normally ,the integral

$$F(x)=\int_0^{\alpha} f(x)dx=\int_0^{\alpha}\alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}dx.$$

Changing the variable $y=x^{\beta}$, we get

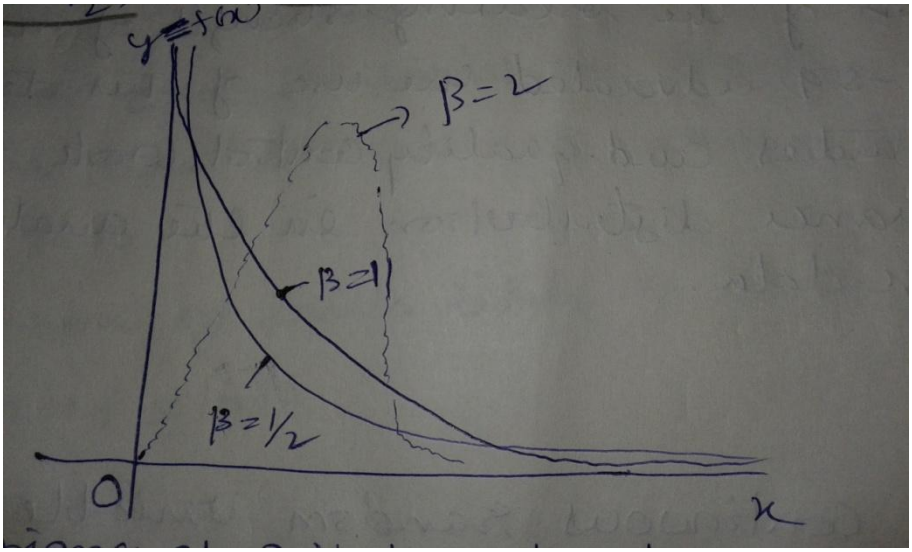
X=0,y=0

$$x=\alpha, y=\alpha^{\beta}, then, dy=\beta.x^{\beta-1}dx$$

$$\therefore\int_0^{\alpha}\alpha.e^{-\alpha y}dy=\alpha.\frac{e^{-\alpha y}}{-\alpha}|_0^{\alpha}$$

$$=\left[e^{-\alpha\alpha^{\beta}}-e^{0}\right]=-\left[e^{-\alpha\alpha^{\beta}}-1\right]=1-e^{-\alpha\alpha^{\beta}}$$

Which is the distribution function of wiebull distribution the graphs of several weibul distribution with $\propto$ =1 and β=1/2,1 and 2 are



Mean and variance of weibull distribution :

The mean of the weibull distribution having the parameters $\propto$ and β may be obtained by evaluating the integral

$$E(X) = \mu = \int_0^\infty x.f(x)dx = \int_0^\infty x.\alpha\beta x^{\beta-1}e^{-\alpha x^\beta}dx$$

Making the change of variable $\mu = \alpha x^\beta$, we get

X=0, u=0, $du = \alpha\beta x^{\beta-1}dx$

$$x = \infty, u = \infty \;\&\; u = \alpha x^\beta \Rightarrow \alpha^{-1}u = x^\beta$$

$$\alpha^{-1/\beta}u^{1/\beta} = x$$

$$\therefore \mu = \int_0^\infty \alpha^{-1/\beta}.U^{1/\beta}.e^{-u}dy$$

$$= \alpha^{-1/\beta}\int_0^\infty U^{\left(1/\beta+1\right)-1}e^{-u}du$$

$$= \alpha^{-1/\beta}\gamma\left(\frac{1}{\beta}+1\right)$$

$\therefore$ Mean of the weibull distribution is $\mu = \alpha^{-1/\beta}\gamma\left(\frac{1}{\beta}+1\right)$ Variance is given by

$$V(x) = E(X^2) - [E(X)]^2 = \int_0^\infty x^2 f(x)dx - \left[\int_0^\infty x.f(x)dx\right]^2$$

i.e., $\int_0^\infty x^2, f(x)dx = \int_0^\infty x^2 \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} dx$

put $\alpha x^\beta = \mu \Rightarrow x = u^{1/\beta}.\alpha^{-1/\beta}, and, x^2 = u^{\frac{2}{\beta}}.\alpha^{-2/\beta}$

x=0, u=0

$x = \infty, u = \infty$

And $du = \alpha\beta.x^{\beta-1}dx$

$$\therefore \int_0^\infty x^2 f(x)dx = \alpha^{\frac{-2}{\beta}} \int_0^\infty U^{\frac{2}{\beta}}.e^{-u}dx$$

$$= \alpha^{\frac{-2}{\beta}} \int_0^\infty U^{\left(\frac{2}{\beta}+1\right)-1} e^{-u}dy$$

$$E(X^2) = \int_0^\infty f(x)dx = \alpha^{\frac{-2}{\beta}}.\gamma\left(1+\frac{2}{\beta}\right)\left(\because e^{-x}x^{n-1}dx = \gamma(n)\right)$$

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

$$= \alpha^{\frac{-2}{\beta}}.\gamma\left(1+\frac{2}{\beta}\right) - \left[\alpha^{\frac{-1}{\beta}}\gamma\left(1+\frac{1}{\beta}\right)\right]^2$$

$$= \alpha^{\frac{-2}{\beta}}\left[\gamma\left(1+\frac{2}{\beta}\right) - \left[\gamma\left(1+\frac{1}{\beta}\right)\right]^2\right]\left(\because E(x) = \mu = \alpha^{\frac{-1}{\beta}}\gamma\left(1+\frac{1}{\beta}\right)\right.$$

Applications :

1. Weibull distribution is widely used to describe the strength distribution of ceramics.
2. In the prediction of the meantime between failures (MTBF)of building components weibull distribution is used.
3. Reliatrility of systems with many components I.e, study of components life times in parallel systems weibull distribution has its application
4. The weibull distribution applied to regional low flow frequency analysis.

Worked out examples :

Example 1: Suppose that the lifetime of am certain kmd ofan emergency backup battery 9in hours0is a random variable `x` having the weibull distribution with $\propto$ =0.1 and β=0.5

Find (i). The mean life time of these batteries

(ii). The probability that such a battery will last more than 300 hurs

Solution : (i). Mean of the weibull distribution is $\mu = \gamma\left(1+\frac{1}{\beta}\right)\alpha^{\frac{-1}{\beta}}$

When $\alpha = 0.1, \beta = 0.5$

$$\mu = \gamma\left(1+\frac{1}{0.5}\right)(0.1)^{\frac{-1}{0.5}}$$

$$= \gamma(3)(0.1)^{-2} = 200hrs.$$

(ii)Performing the necessary integration we get

$$\int_{300}^{\infty} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}dx = \int_{300}^{\infty}(0.1)(0.5)x^{-0.5}, e^{-0.1x^{0.5}}dx$$

$$\int_{300}^{\infty}(0.05)\frac{1}{\sqrt{x}}e^{-0.1\sqrt{x}}dx$$

$$= 0.05\int_{300}^{\infty} x^{\frac{-1}{2}}e^{-0.1x^{\frac{1}{2}}}dx$$

$$= e^{-0.1(300)^{0.5}}$$

$$= 0.177$$

$\therefore$ The probability that such a battery will last more than 300 hrs is 0.177.

Reliability :

Definition : Reliability of a product or item is defined as the probability that it will function within specified limits for at least a specified period of time under specified environmental conditions .then the probability that the component will fail on the interval fom 0 to t is given by`

$$P(T \le t) = F(t) = \int_{0}^{t} f(x)dx$$

And the reliability function expression the probability that it survives to time t ,is given by

$$P(T > t) = R(t) = 1 - F(t)$$

And failure rate function is given by

$$Z(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)}$$

Exponential distribution plays an important part in life testing and reliability problems for a situation where the failure rate appears to be more or less constant , the exponential distribution would be an adequate choice but not all items satisfy the condition that it does not age in the life lasting research the simplest and the most widely exploited model is the one parameter exponential distribution with p.d.f.

$$f(x) = \theta.\exp(-\theta x), x \geq 0, \theta > 0$$

Then distribution function is $F(t) = 1 - \exp(-\theta t), t > 0, \theta > 0$ & Reliability function is $R(t) = 1 - F(t) = ex[(-\theta t)$

Failure – rate function $Z(t) = \frac{F(t)}{R(t)} = \frac{\theta \exp(-\theta t)}{\exp(-\theta t)}$

Normal Distribution plays a very important role in statistical theory as well as methods the normal distribution also arises as a limiting from of various other distributions. In the content of life testing and reliatlety problems the normal distribution give quite a good fit for the failure time data the p.d.f. of the normal distribution with parameters μ and σ is given by

$$f(x/\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\left[\exp - \left\{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\right] - \infty < x < \infty, -\infty < x < \infty$$

The corresponding distribution function F(x, μ, σ) is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp - \left(\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) du = \phi\left(\frac{x-\mu}{\sigma}\right)$$

Then the reliability function is given by

$$R(x) = 1 - F(x) = 1 - \phi\left(\frac{x-\mu}{\sigma}\right)$$

Weibull distribution adequately describes the failure times of components when their failure rate either increases or decreases with time. It has the parameters α and β then its p. d. f. is given by

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}, t > 0, \alpha > 0, \beta > 0$$

Then the corresponding distribution function of weibull distribution is

$$F(t) = 1 - e^{-\alpha t^{\beta}}.$$

Then reliability function is given by

$$R(t) = 1 - F(t) = 1 - (1 - e^{-\alpha t^{\beta}}) = e^{-\alpha t^{\beta}}$$

And the weibull failure rate function is given by

$$Z(t) = \frac{f(t)}{R(t)} = \frac{\alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}}{e^{-\alpha t^{\beta}}} = \alpha\beta t^{\beta-1}$$

Worked out examples :

Example 1 : 50,000 units of product X was sent to the field for sales,After 6 months it was determined that 156 failures had occurred with an average estimated operating time of 730 hrs .How many of these units will fail in the first 3000 operating hours ?

Solution : Let " θ "be the average failure time and is obtained as

$$\theta = \frac{Totalfails}{totaloperatingtime} = \frac{156}{50{,}000 \times 730} = 4.27 \times 10^{-6}\ failures / hour$$

Since X follows an exponential distribution , the reliability at 3000 hours is obtained as

$$R(t) = \exp(-\theta t) = \exp(-4.27 \times 10^{-6} \times 3000) = 0.987$$

These ,the number of units survivor after 3000 hours is 50,000×0.0987=49,350 and number of units failed before 3000 hours is 650 units

Example 2: A certain type of electric components has a uniform failure rate of 0.00001 per hour .what is its reliability for a specified period of service of 10,000 hrs/

Solution : Given θ =0.00001per hour

t=10,000 hrs

if the random variable follows an exponential distribution then

$$R(t) = e^{-\theta t} = e^{-0.00001 \times 10{,}000}$$

$$= e^{-0.1}$$

$$= 0.90483, or, 90.483, percent$$

the reliability for a specified period of 10,000 hrs is 0.90483

Example 3 : Given λ =5000 hrs and uniform failure rate , what is the reliability associated with a specified since period of 200hrs ?

Solution : Given λ =5000 hr

t =200hrs

Now $\theta = \frac{1}{\lambda} = failure, rate = \frac{1}{5000}$

$$R(t) = e^{-\theta t}$$

$= e^{\frac{-1}{5000}\times 200}$

$= e^{-0.04}$

$= 0.96079, or, 96.079, percent$

Her the reliability associated with a specified service period of 200 hrs .is 0.96079.

1.8. Excercises:

1. Two cards are drawn at random from a well shuffled pack of cards show that the probability of drawing two aces is 1/221.

2. Among the digits 1,2,3,4,5 at first one is chosen and then a second selection is made among the remaining from digits . Assuming that all 20 possible out comes have equal probabilities find the probability that an odd digit will be selected

(i) The first time Ans : 3/5

(ii) The second time Ans : 3/5

(iii) Both the times. Ans : 3/10.

3. out of (2n+1) tickets consecutively numbered three are drawn at random. Find the chance that the numbers are in A.P

Ans: $\frac{3n}{4n^2-1}$

4. If two dice are thrown what is the probability that assumes

(a) greater than 8

(b) neither 7 nor 11.

Ans : 5/18, b) 7/9.

5. A box contains 6 red , 4 white and 5 black balls a person draws 4 balls from the box at random find the probability that among the balls drawn there is at least one ball of each color

Ans: 0.5275

6. Each coefficient of the equation $ax^2 + bx + c = 0$ is determined by throwing an ordinary due. Find the probability that the equation will have real roots.

Ans : 43/216

7 . if $A \bigcap B = \phi$ then show that $P(A) \le P(B)$

8 . IF A and B are two events such that

$P(A) = \frac{3}{4}, P(B) = \frac{5}{8}, shwthat$

$P(A \cup B) \ge \frac{3}{4}, \frac{3}{8} \le P(A \cap B) \le \frac{5}{8}$

9. A special dice is prepared such that the probabilities of throwing 1,2,3,4,5,6 are respectively

$\frac{1-k}{6}, \frac{1+2k}{6}, \frac{1-k}{6}, \frac{1+k}{6}, \frac{1-2k}{6}, \frac{1+k}{6}$, respectively

If two such dice are thrown find the probability of getting a sum equal to 9.

1. A consignment of 15 record players contains 4 defectives. The record players are selected at random, one by one and examined. Those examined are not put back. What is the probability that the 9^{th} one examined is the last defective ?

$\frac{8}{195}$

2. it is given that $P(A_1 \cup A_2) = \frac{5}{6}, P(A_1 \cap A_2) = \frac{1}{3}, ANDP(A_2) = \frac{1}{2}, WHEREP(\bar{A}_2)$ stands for the probability that A_2 does not happen. Determine $p(A_1)$ and $p(A_2)$ and hence show that A_1 and A_2 are independent.

Ans $p(A_1)=2/3$, $p(A_2)=1/2$ and A_1, A_2 are independent.

3. 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 draw 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.

Ans: $\frac{6_{c_4}}{15_{c_4}} X \frac{9_{c_4}}{15_{c_4}}$

$\frac{6_{c_4}}{15_{c_4}} X \frac{9_{c_4}}{11_{c_4}}$

4. The chances that doctor A will diagonose a disease X correctly is 60% the chances that a patient will die by his treatment after connect diagnosis is 40% and the chance of death by wrong diagnosis is 70% . A patient of doctor A, who had desease X, died. What is the chance that his desease was diagnosed correctly?

6/13 answer.

5. The contents of units I,II, and III are as follows

Unit I : 1 white , 2 black and 3 red balls

Unit II : 2 white, 1 black and 1 red balls, and

Unit III : 4 white , 5 black and 3 red balls.

One urn is chosen at random and two balls drawn. They happen to be white and re. what is the probability that they come form urus I, II, or III ?

Answer : $\frac{33}{118}, \frac{55}{118}, \frac{30}{118}$

6. In a bolt factory machines A, B, and C manufacture respectively 25%, 35% and 40% of the total. If their output 5,4,2 percent are defective bolts . A bolt is drawn at random form the product and it is found to be defective. What are the probabilities that it was manufactured by machines A, B and C ?

Answer : $\frac{25}{69}, \frac{28}{69}, \frac{16}{69}$

7.There are 12 cards numbered 1 to 12 in box. If two cars are selected what is the probability that sum is odd. (i) with replacement (ii) with out replacement.

Ans ¼,6/11.

8. The odds against A solving a certain problem are 4 to 3 and odds infacour of B solving the same problem are 7 to 5. What is the probability that the problem is solved if they both try independently ?

Answer 16/21.

9. In answering a question on a multiple choice test a student either rows the answer or he guesses .let p be the probability that he rows the answer and 1-p the probability that be quessel. Assume that a student who guesses at the answer will be correct with probability 1/5, where 5 is the number of multiple choice alternatives. What is the conditional probability that a student know the answer to a question given that he answered it correctly?

Answer : $\frac{5p}{4p+1}$

10. In a certain town 40% have blown hair, 25% have brown eyes and 15% have bothe brown hair and brown eyes. A person is selected at random form the town

(i) if he has brown hair, what is the probability that he has brown eyes also.

(ii) If he has brown eyes, determine the probability that he does ot have brown hair.

Answer (i) 0.375

(ii) 0.6

11. Companies B_1,B_2, B_3 produce 30% , 45% and 25% of the cars respectively. If is known that 2%. 3% and 2% of these cars produced form B_1, B_2,B_3 are defective.

(i) what is the probability that a car punchased is defective.

(ii) if a car purchased is found to be defective what is the probability that this car produced by the company B.

Answer : 0.0245

(ii) 0.2449

1. The diameter of an electric cable say X, is assumed to be a continuous random variable with p.d.f f(x)=6x(1-x), $0 \le x \le 1$,

(i) Check that above is pdf

(ii) Determine a number such that p(x<b)=p(x>b)

Ans (i) pdf

Ans (ii) b=1/2.

2. A continuous random variable has the pdf $f(x) = kxe^{-\lambda x}$ if $x \ge 0, \lambda \ge 0$ and 0 other wise . Determine the constant k, find mean and variance.

Ans : $k = \lambda^2, mean = \frac{2}{\lambda}, \text{var}\, iances = \frac{2}{\lambda^2}$

3. Find the m.g.f of the random variable 'X' having pdf

$f(x) = \{x, for 0 \le x < 1$
$2 - x, for, 1 \le x < 2$
$0, otherwise$

Ans : $M_x(t) = \frac{(e^t - 1)^2}{t^2}$

4. Tow dice are thrown X assign to each point if s the sum of the variables on the faces . Find mean and variances of the random variable.

Ans : Mean $\mu = 7, \text{var}\, ience\, \sigma^2 = 5.8$

5. A fair coin is tossed until a head or five tails occur . Find the expected number of E of tosses of the coin.

Ans : $E(x) = \mu = 1.9$

6. Calculate expectation and variance of x. if the probability

X	-1	0	1	2	3
P(x)	0.3	0.1	0.1	0.3	0.2

Ans : E(x)=1, V(x)=2.4

7. A petrol pump is supplied with petrol once a day .if its dialy volume x of sales in thousands of letres is distributed by

F(x) = $5(1-4)^4; 0 \le x \le 1$

What must be the capacity of its tank in order that the probability that its supply will be exhausted in a given day shall be 0.01?

Ans : a=0.6019, 601.9 liters.

8. verify that the following function is a distribution

$$f(x)=\begin{cases}0, x<-a\\ \frac{1}{2}(\frac{x}{a}+1) -a\le x\le a\\ 1, x>a\end{cases}$$

9. let $f(x)=\begin{cases}\frac{1}{2}, -1<x<1\\ 0, otherwise\end{cases}$

Be the pdf of the r.v.X. Find distribution function and the p.d.f of $y=x^2$?

1. Suppose the life of automobile batteries is exponentially distributed with parameter θ=0.001 days

   What is the probability that a battery will last more than 1200 days ?

   Ans : 0.301

2. The life time X in hours of a T.V. table of certain type obeys an exponential distribution with θ=0.001 hrs
   Find (i)P(x>1000) (ii)P(700≤×≤1000)
   Ans: (i) 0.368, (ii) 0.129
3. A continuous random variable X has the probability density function f(x) given by

$$f(x)\begin{cases}A.e^{\frac{-x}{5}}, x>0\\ 0, otherwise\end{cases}$$

   Find the value of A

4. Suppose that during raining season on a tropical island the length of the shower has an exponential distribution with parameter θ-2,time being measured in minutes what is the probability that a shower will last more than three minutes ?if a shower has already lasted for 2 minutes ,what is the probability that it will last for atleast one more mite ?
5. Find the probability that a random variable having the standard normal distribution will take on a value between 0.87 and 1.287

   Ans: 0.0919
6. If x is a normal variate with mean 30 and S.D.S.
   Find the probability that (i)26≤×≤40 (ii) x≥45.

Ans: (i) 0.7653, (ii) 0.0013

7. In a distribution exactly normal 7% of the intense are under 35 and 89% are under 63.what are the mean and S.D of the distribution .
   Ans : $\mu = 50.3, \sigma = 10.3$

8. 1000 students had written an examination the mean of test is 35 and S.D IS 5. Assuming the distribution to be normal find

   (i). how many students marks lie between 25 and 40
   (ii). how many students get more than 40
   (iii).how many students get below 20
   (iv). How many get more than 50

   Ans : (i) 819, (ii) 159, (iii) 1, (iv) 1

9. It has been claimed that in 60% of all solar heat installation the utility bill is rescued at least one third .Accordingly what are the probabilities that the utility bill be reduced by at least one

(i). from of the five installation

(ii). at least four of five installations

Ans : (i) 0.259, (ii) 0.337

10. An insurance agent accepts policies of 5 men all of identical age and in good health. The probability that a man of this age will be alive 30 years is 2/3 find the probability that in 30 years (i). all five man (ii).at least one man (iii).at most three will be alien

    Ans : (i) $\frac{32}{243}, (ii), \frac{242}{243}, (iii)\frac{131}{243}$

11. Using Poisson distribution . Find the probability that an ace will be drawn from a pack of 52 cards is exactly once among 52 consecutive trials
    Ans: 0.368

12. If the variance of a Poisson vitiate is 3. Find the Probability that (i)x=0 (ii)1 $\leq x < 4$ (iii) $0 < x \leq 3$

    Ans : (i) 0.0498, (ii) 0.5976, (iii) 0.5976

13. A manufacturer of pins knows that 2% of this product is defective .if he sells pins in bones of 100 and guarantees that not more than 4pins will be defective what is the probability that a box will fail to meet the guaranteed quantity

    Ans : 0.048

14. Define simulation and explain the simulation of a discrete random variable with an example

15. Explain the concept of reliability and given the expression for reliability and given the expression for reliability in case of exponential and weibull distribution

19. Summary

The concept probability is defined in various ways starting from the classica definition to the most modern way the aniomate approach. Some general laws of probability upto the concept of adective law for two and more than two events are established besides showing the applications of these laws in a number of examples. Some excersises in the answers one also provided for the students to try on their own.

In this lesson an attempt is made to explain the concepts of conditional probability and related aspects along with examples .the most important aspect is the Baye's theorem and inverse probabilities A number of examples are worked out and a good number of exercises are also given.

The concept of random variable , its associated distribution function are defined and a number of examples are given for both discrete and continuous r is the associated mass functions and density functions and also mathematical expectation and moment generating function are presented

1.10. Technical terms.

Relative frequency, random experiment, sample space , single event, compound event, anions.

Dependent Events

Prior probability

Posterior/ inverse probability

Mathematical induction

Product law

Independent Events

Mutual independence

Relative probability

Pairwise independence.

Random variable, distribution function probability mass function probability density function , mathematical expectation, moment generating function

UNIT II

ESTIMATION AND TEST OF HYPOTHESIS -I

SYLABUS

Point estimation, interval estimation, central limit theorem Hypothesis testing and significance test on the mean and variance estimating one and two proportions, testing hypothesis of a proportion and two proportion.

Objective :

After studying this unit the student is expected to have a clear idea about estimation (point, interval) and tests of hypothesis and their applicabilities.

Structure of the Unit:

2.1 Introduction

2.2 Estimation

(i) Point estimation

(ii) Interval estimation

(iii) central limit theorem

2.3 inferences on the mean and variance of a distribution

(i) Hypothesis testing.

(ii) Significance testing.

(iii) Hypothesis and significances test on the mean.

(iv)hypothesis tests on the variable.

2.4 inference on proportion

(i) estimating proportion

(ii) testing hypothesis on a proportion

(iii) estimating two proportion

(iv) testing hypothesis of two proportions.

2.5 worked out examples

2.6 exercise

2.7 summary

2.8 technical terms.

2.1 Introduction:

Statistical inference is the process of drawing various conclusions concerning population characteristics called parameters from the sample values. Estimation about the various parametric value such as

$\mu or \sigma^2$ etc or the determination whether an observed difference between two sample estimates of a parametric value is just by chance or it is because they were drawn from different populations and are significantly different from one another etc. are the problems of inference. The problem of estimation which was found by prof. R.A.Fisher through his series of fundamental papers in1930's.

After discussing the point estimation and setting up the confidence internals for parameters we come across the problem of testing the statistical hypothesis. Suppose some business concern has an average sale of Rs. 5000/- daily estimated over a long period . A new sales girl claims that she will increase the average sales by Rs. 300/- a day. The concern is interested in an increased sale no doubt, but how to know whether the claim of the girl is justified or not?for this some such a mathematical model for the population of increased sales is assumed which it agrees to the maximum with the practical observations. In the example given, let us assume that the claim of the girl about her sales is justified and that the increase in sales is normally distributed with mean $\mu = 300$ and variance σ^2. This assumption is called statistical hypothesis. Thereafter the suitability of the assumed model is examined on the basis of the dale observations made. This procedures called testing of hypothesis.

In actual practice the distribution of sample values would not have an absolute similarity with the assumed population distribution. So generally we are to depend on the degree of closeness between the two. The procedure of measuring the closeness between the assumed model and the observed phenomenon is an for testing of hypothesis and which in itself is called testing of significances.

2.2 Estimation.

(i) Point Estimation.

Estimate definition:

To find an unknown population parameter, a judgment or statement is made which is an estimate.

Estimator definition:

The method or rule to determine an unknown population parameter is called an estimator . For example , sample mean is an estimator of population mean because sample mean is a method of determining the population mean. Let x1, x_2,……X_n, is a sample taken from a population whose probability density function is f(x, t) where t is the known parameter. A parameter can have many or (1,2) estimations the estimators should be found so that they are very nearer to the parameter values. These are two types of estimation

(i) Point estimation
(ii) Interval estimation

Point Estimation:

If from the observation, in sample , a single value is calculated as an estimate from an unknown population parameter the procedure is referred to as point estimation.

Properties of good Estimator:

An estimator is said to be a good estimator is it is

(i) Unbiased
(ii) Consistent
(iii) Efficient
(iv) Sufficient unbiased estimator

A statistic $t=t(x_1,x_2,\ldots X_n)$ a function of the sample values $X_1,X_2, \ldots x_n$

is an unbiased estimator of population parameter θ, if $E(t) = \theta$. In other words is

E(statistic)=parameter

Suppose if $E(t) > \theta$ then, t is called positively biased and if $E(t) < \theta$, then , t is called negatively based

Example:

Let $x_1,x_2,\ldots\ldots x_n$ be a random sample drawn from a given population with mean μ and variance σ^2 Show that sample mean $\bar{x}$ is an unbiased estimate of population mean μ i.e., $E(\bar{x}) = \mu$.

Solution :

Let $x_1,x_2,\ldots\ldots x_n$ be a random sample drawn from a given population with mean μ and variance σ^2 then mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{---------- (1)}$$

Taking expectation on both sides of (1)

$$E(\bar{x}) = E[\frac{1}{n}\sum_{i=1}^{n} x_i]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}E(x_1 + x_2 + \ldots\ldots + x_n)$$

$$= \frac{1}{n}\left[E(x_1) + E(x_2) + \ldots\ldots + E(x_n)\right]$$

$\because E(x_i) = \mu, i = 1,2,\ldots n$

We have $E(\bar{x}) = \frac{1}{n}[\mu + \mu + ... + \mu] = \frac{1}{n}.n\mu = \mu$

$\therefore \bar{x}$ is an unbiased estimator μ.

Example : Let $x_1, x_2, \ldots x_n$ be a random sample of size n drawn from a finite population then $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is not unbiased estimate of the parameter σ^2 but $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is unbiased.

Solution: $x_1, x_2, \ldots x_n$ be a random sample of size n drawn from a finite population then $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ after taking expectation on both sides we have

$$E(S^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n}E\left[\sum_{i=}^{n}(x_i - \mu + \mu - \bar{x})^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}[(x_i - \mu) - (\bar{x} - \mu)]^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{2}{n}(x_i - \mu)(\bar{x} - \mu) + \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - \mu)^2\right]$$

$$= \left[\sigma^2 - E\left[2(\bar{x} - \mu)^2\right] + (\bar{x} - \mu)\right]$$

$$\because \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{n}{n}\mu = \bar{x} - \mu$$

Also $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{n}n(\bar{x} - \mu)^2 = (\bar{x} - \mu)^2$

$$= \sigma^2 - E(\bar{x} - \mu)^2 = \sigma^2 - \frac{\sigma^2}{n} = (\frac{n-1}{n})\sigma^2$$

$\therefore E(S^2) \neq \sigma^2$

$\therefore \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is not an unbiased estimator of σ^2

Now we write $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$E(S^2) = E\left[\frac{n}{n-1}\sum_{i=1}^{n}\frac{1}{n}(x_i - \bar{x})^2\right]$$

$$= E\left[\frac{n}{n-1}S^2\right] = \frac{n}{n-1}E(s^2) = \frac{n}{n-1}\left(\frac{n-1}{n}\right)\sigma^2$$

$$\because \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$$

$$\therefore E(S^2) = \sigma^2$$

Hence $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator of σ^2 .

More efficient unbiased estimator:

If $\bar{x}_1, \bar{x}_2$, are two unbiased estimators of $\mu, and, \sigma_1^2, \sigma_2^2$ are variances of their sampling distributions and $\sigma_1^2 < \sigma_2^2$ then $\bar{x}_1$ is said to be more efficient unbiased estimator of μ.

Note: if $\bar{x}_1, \bar{x}_2 ..... \bar{x}_k$ are unbiased estimators of the parameter μ, $\sigma_1^2, \sigma_2^2 ..... \sigma_k^2$ are variances of their sampling distribution then $\bar{x}_1$ is the most efficient estimator of μ if $\sigma_1^2 < \sigma_2^2 < ..... < \sigma_k^2$.

i.e., e= efficiency of an estimator of $\bar{x}_k$

$$\Rightarrow e = \frac{\text{var}\,\bar{x}_1}{\text{var}\,\bar{x}_i}, i = 2,3,...k.$$

e<1

Maximum error of estimate:

For large n Z= $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is a random variable having approximately the standard normal distribution. with probability 1-α the inequality

$$-Z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}$$ will be satisfied

$$i.e., \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}$$

When $Z_{\alpha/2}$ is such that the normal curve area to its right equal to $\alpha/2$. Let Ebe the maximum error of $|\bar{x}-\mu|$ that is, maximum error of estimates. Hence maximum error of estimates E= $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. with probability 1-$\alpha$ which means we can assert with probability 1-α that the error $|\bar{x}-\mu|$ will be at most $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Note: the most widely used values for 1-α are 0.95 and 0.99 and the corresponding values of $Z_{\alpha/2}$ are $Z_{0.025}$=1.96 and $Z_{0.05}$=2.575.

Determination of sample size:

We know that maximum error E= $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

$$\therefore \frac{Z_{\alpha/2}.\sigma}{E}=\sqrt{n}$$

Hence $n=\therefore\left[\frac{Z_{\alpha/2}.\sigma}{E}\right]^2$

Sample size n = $\left[\frac{Z_{\alpha/2}.\sigma}{E}\right]^2$

Where σ is standard deviation of population

E is the minimum error.

$Z_{\alpha/2}$ is the critical value of Z.

Maximum error estimate for small samples:

When sample size is small we know that $t=\frac{\bar{X}-\mu}{S/\sqrt{n}}$ then with probability 1-α we have $\frac{\bar{X}-\mu}{S/\sqrt{n}}<Z_{\alpha/2}$.

As similar to that of large sample, the maximum error estimate E is

$$E=t_{\alpha/2}.\frac{s}{\sqrt{n}}$$

Determination of sample size:

From maximum error estimate E for small samples the sample size is given by

$$\sqrt{n}=\frac{t_{\alpha/2}.s}{E}$$

i.e., n = $\left[\frac{t_{\alpha/2}.s}{E}\right]^2$

Where E is maximum error estimate

S is standard deviation of sample.

N is the sample size.

$t_{\alpha/2}$. is the critical value of t.

Interval Estimation:

The point estimate does not in general, coincide with the true value of the parameter. If is thus preferred sometimes to obtain a range or the interval of values which may be expected to cover the true value of the parameter with some defined probability or the degree of confidence. Such an interval is called the interval estimate or the confidence interval and the probability or degree of confidence is called the confidence coefficient. Even through different confidence coefficient would.give different confidence intervals, so when defining a confidence interval it must be accompaniedby the confidence coefficient.

Let xi(i=1,2,…n) be a random sample of n Observations from a population involving a single unknown parameter θ says.

Let f(xi, θ) be the probability function of the parent distribution from which the sample is drawn and let us suppose that this distribution is continuous. Let $t = t(x_1, x_2, ...x_n)$ a function of the sample values be an estimate of the population parameter θ , with the sampling distribution given by g(t, θ)

Having obtained the value of the statistic 't' from a given sample, the problem is "can we make some reasonable probability statements about the unknown parameter θ in the population from which the sample has been drawn ?" this is very well answered by the techniques of confidences internal due to Nyman and is obtained below.

For all some small value of α (5% or 17) and then determine two constants say c1, and c2 such that

$$p(c_1 < \theta < c_2) = 1 - \alpha$$

The quantities c1 and c2 so determined, are known as the confidence limits and the interval (c1,c2) with in which the unknown value of the population parameter is expected to lie called the confidence interval and (1- α) is called the confidence coefficient.

Thus if we take α =0..5(or 0.01), we shall get 95% or 99% confidence limits. Let T1 and T2 be two statistics such that

$$p(T_1 > \theta) = \alpha_1$$

$$p(T_2 < \theta) = \alpha_2$$

Where $\alpha_1 and \alpha_2$ are constants independent of equations 2 and 3 combined to sries

$p(T_1 < \theta < T_2) = 1 - \alpha$

Where $\alpha = \alpha_1 + \alpha_2$

Confidence interval for mean:

Let μ and σ be the mean and standard deviation of a normal population and

$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ ~ N(0,1)

Which means Z is standard normal variant with o mean and unit standard deviation. If $Z_{\alpha/2}$ is a value of Z such that area under the normal curve then the probability that Z lies between $-z_{r/_2}$ to $z_{r/_2}$

Is (1-α) or

$$i.e., P\left[-Z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right] = 1 - \alpha$$

$$\therefore P\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

Then $\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ and $\left[\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ is the (1-α) 100% confidence interval for the population mean

Now we can say that with (1-α) 100% confidence that the interval from

$\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ to $\left[\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ contains μ.

$\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right], \left[\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ is called the confidence interval for μ. having the degree of confidence (1-α) 100% hence $\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ and $\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ are called confidence limits.

Confidence interval when σ is unknown (small samples) suppose a random sample of size n taken from a normal population then t= $\frac{\bar{x} - \mu}{s/\sqrt{n}}$

Where $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{n-1}$, is a random variable with v=n-1 d.f

$$P(-t_{\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{\alpha/2}) = 1-\alpha$$

$$P(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$$

(1-α) 100% confidence interval of μ when σ is unknown is

$$(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}})$$

Hence $\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}$, and $\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$ are confidence limits for small samples

($\sigma unknown$)

Central limit theorem:

The sampling distribution of the mean ($\sigma - known$)

Suppose same statistics 't' for ex: x or r^2etc is computed from all N_{cn} samples of size n taken from same specified population of size N. we will find that the value of t would vary from one sample to another to the extent that from a large or an infinite population we shall have a series of z values . These values can be grouped into the form of a frequency distribution this is known as the sampling distribution of the statistic.

That is, if we draw a sample of size n form a given finite population of size N, then the total number of possible samples is

$$N_{C_n} = \frac{N!}{n!(N-n)!} = k(say)$$

For each of these K samples we can compute some statistics $t = t(x_1, x_2, ...X_n)$ in particular the mean $\bar{x}$, the variance s^2 etc as given below.

Sample number	Statistics T $\bar{x}$ s^2
1 2 3 . . . K	$T_1\ \bar{x}_1\ s_1^2$ $T_2\ \bar{x}_2\ s_2^2$ $T_3\ \bar{x}_3\ s_3^2$ . . . . $T_k\ \bar{x}_k\ s_k^2$

The set of the values of the statistic so obtained, one for each sample, constitutes what is called the sampling distribution of the statistic. For example, the values $t_1, t_2, t_3, \ldots t_k$ determine the sampling distribution of the statistic t. In other words, statistic t may be regarded as a random variable which can take the values $t_1, t_2, t_3, \ldots t_k$ and we can compute

the various statistical constants like mean, variance , skewness, kurtosis etc. for its distribution for example the mean and variance of the sampling distribution of the statistic t are given by

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + ... + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i$$

$$\text{Var(t)} = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + ... + (t_k - \bar{t})^2\right] = \frac{1}{k}\sum_{i=1}^{n}(t_i - \bar{t})^2$$

For example :

Consider the collection of 25 samples of size 2 taken from a population of size 5 with replacement. With replacement we can select 5x5=25 samples.

Population =(1,2,3,4,5,)

The mean of the population = $\mu = \frac{1+2+3+4+5}{5} = 3$

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 sampling distribution.

(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)

Mean of the sampling distribution of mean

$$\mu_{\bar{x}} = \frac{1+1.5+2+2.5+.....+4.5+5}{25} = 3$$

Mean of these 25 sample are

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

$\mu = \mu_{\bar{x}} = 3$

Similarly we can calculate the variance

$\sigma_{\bar{x}}^2 = \sum_i \frac{(x_i - \mu_{\bar{x}})^2}{n} = 1$, so the $\sigma^2 \neq \sigma_{\bar{x}}^2$

To find the mean of the sampling distribution of means:

Let $\bar{x}_1 + \bar{x}_2, ...be$ the means of these samples, μ, and, σ^2 be mean and variance of finite population of size N.

Let $\bar{x}$ be mean of sample of size n and if we take number of sample of size n.

$$E(\bar{x}) = \frac{E(x_1 + x_2 + ... + x_{n)}}{n} = \frac{1}{n} E(x_1 + x_2 + ... + x_n)$$

$$= \frac{1}{n}\left[E(x_1) + E(x_2) + ... + E(x_{n)}\right]$$

Since the sample is random we can assume that

$$E(x_1) = E(x_2) = E(x_3) = ...E(x_n) = \mu$$

$$\therefore E(\bar{x}) = \frac{1}{n}(\mu + \mu + ...\mu times) = \frac{n\mu}{n} = \mu$$

The mean of sampling distribution of means = the mean of finite population of size N.

i.e, $\mu = \mu_{\bar{x}}$

To find the variance of sampling distribution of means suppose σ^2 is the variance of population.

Let $v(\bar{x}) = \sigma_{\bar{x}}^2$ be the variance of sampling distribution of means then

$$\sigma_{\bar{x}}^2 = \text{var}(\bar{x}) = \text{var}(\frac{1}{n}\sum_{i=1}^{n} x_i) = v\left[\frac{x_1 + x_2 + ...x_n}{n}\right]$$

$$= \frac{1}{n^2} V(x_1 + x_2 + ...x_n)$$

$$= \frac{1}{n^2}(v(x_1) + v(x_2) + ... + v(x_n)$$

Since the sample is random, we can assume that

$$v(x_1) = v(x_2) = ... = v(x_n) = \sigma^2$$

$$\text{var}(\bar{x}) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + ... + \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\therefore \sigma_{\bar{x}}^2 \neq \sigma^2, \sigma_{\bar{x}}^2 = \frac{\sigma^2}{\sqrt{n}} \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Hence the variance of sampling distribution of means $\neq$ the variance of population.

Note :1 As n increases, the distribution of $\bar{x}$ is more concentrated about the mean μ.

2. The formula $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{\sqrt{n}}$ is applicable if the sampling is simple and with replacement.

3. if the sampling is without replacement $\sigma_{\bar{x}}^2 = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$

Where N is the size of the population and n is the sample size and $n \leq N$. Also $\left(\frac{N-n}{N-1}\right)$ is called the finite population connection factor.

Standard error (S.E)

The standard deviation of the sampling distribution of a statistic is known as standard error.

$$\therefore S.E(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Note:1 Standard error plays a very important role in the theory of large samples.

2. if t is any statistic , then for large samples $Z = \frac{t - E(t)}{\sqrt{\text{var}(t)}} \sim N(0,1)$ asymptotically as $n \to \infty$

Hence $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$ asymptotically as $n \to \infty$.

As we know that if n is large i.e., as $n \to \infty$ the binomial distribution tends to normal , so that in the case of large samples, properties of normal distribution can be applied. Suppose we want to test the hypothesis that a given large sample of size n is obtained by simple sampling from a population for which the probability of success is p. Since for normal distribution the probability that a random variable lie between $(-3\sigma, 3\sigma)$ is 0.9973 Mean +-3.s.e mean .But for binomial distribution mean =np and S.E= $\sqrt{npq}$

Np+-3. $\sqrt{npq}$, which means 99.73% are within the range and 3% are outside the range.

Central limit theorem:

It is impossible to determine exact form of the distribution without the knowledge of the actual form of the population, but it jis possible to find the limiting distribution as $n \to \infty$ of a random variable whose, values are closely related to $\bar{x}$, assuming only that the population has a finite variance σ^2. The random variable here concerned here is the standardized sample mean.

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

Whose values are given by the differences between $\bar{x}$ and μ

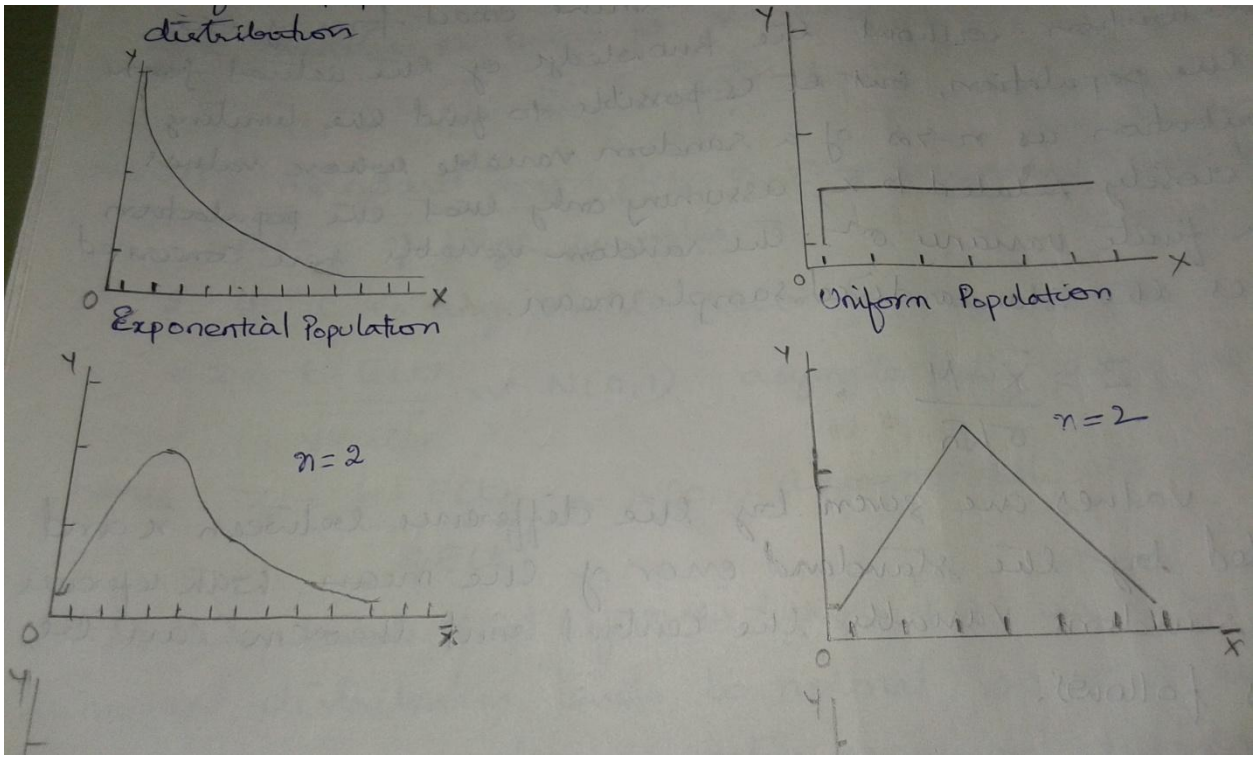
Divided by the standard error of the mean. With reference to this random variable the central limit theorem can be stated as follows.

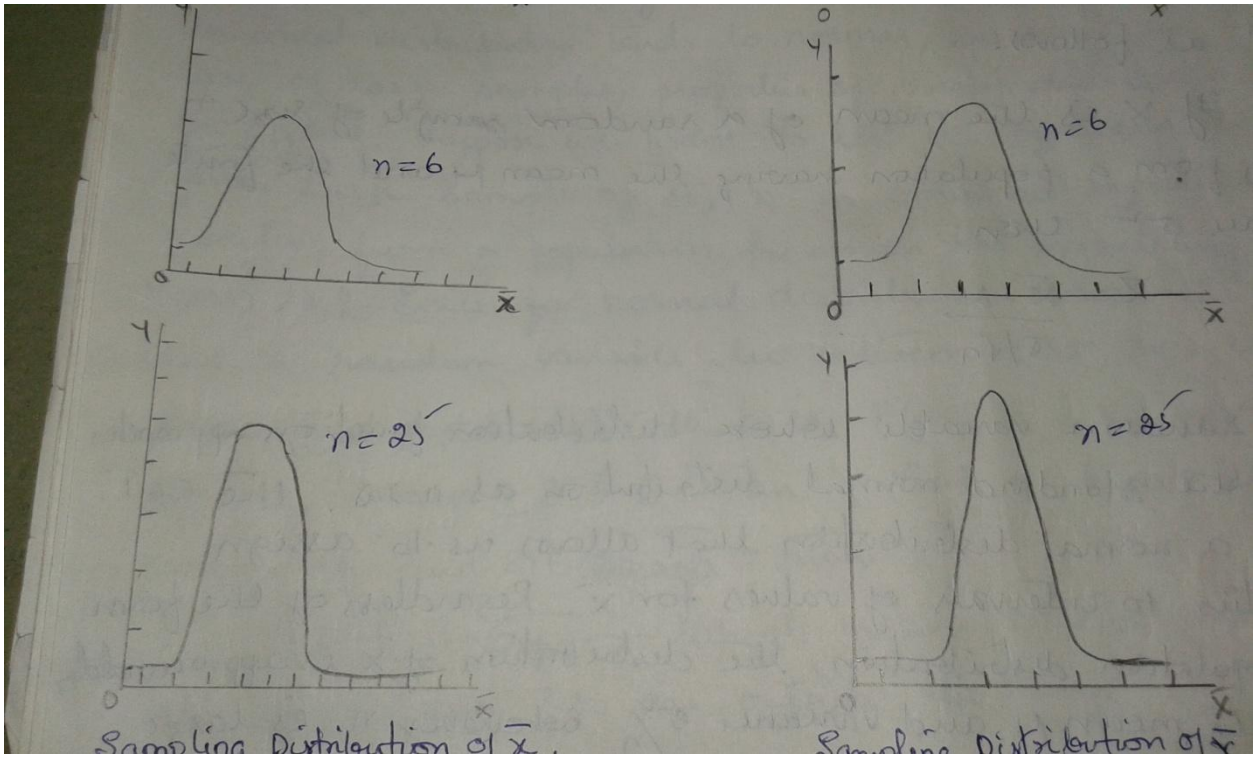
If $\bar{x}$ is the mean of a random sample of size n taken from a population having the mean μ and the finite variance σ^2 then

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

Is a random variable whose distribution function approaches that of the standard normal distribution as $n \to \infty$. The CLT provides a normal distribution that allows us to assign probability to intervals of values for $\bar{x}$. Regardless of the form of the population distribution, the distribution of $\bar{x}$ is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$

Whenever n is large. This tendency towards normality is illustrated for uniform population distribution and an exponential population distributions.





In practice, the normal distribution provides an excellent approximation to the sampling distribution of the mean for n as small as 25 or 30. As we saw in figures of exponential and uniform population the sampling distribution of the mean has the general shape of a normal distribution even for samples of size x=10 from a discrete uniform distribution . if the random samples come from a normal population, the sampling distribution of the mean is normal regardless of the size of the sample

Example:

If a one gallon paint covers on the average 513.3 square feet with a standard deviation of 31.5 square feet, what is the probability that the mean area covered by a sample of 40 of these one gallon cans will be anywhere from 510.0 to 520.0 square fact?

Solution: Given one gallon paint covers on the average $\mu = 513.3$ square feet with standard deviation $\sigma = 31.5 sqft$.

Here we have to find probability that the mean area covered by a sample of 40 as these one gallon cans will be anywhere from 510.0 to 520.0 which means we have to find normal curve area between

$$Z = \frac{510.0 - 513.0}{31.5/\sqrt{40}} = 0.66 and Z = \frac{520 - 513.3}{31.5/\sqrt{40}}$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

From the normal tables we obtain a probability of 0.6553. Note that if $\bar{x}$ turned out to be much less than 513.3, say, less than 500.0, this might create a doubt whether the sample actually come from a population having $\mu = 513.3 and \sigma = 31.5.$ the probability of obtained such a small value i.e. a z value less than -2.6r is only 0.0038.

The sampling distribution of the mean (σ Unknown):

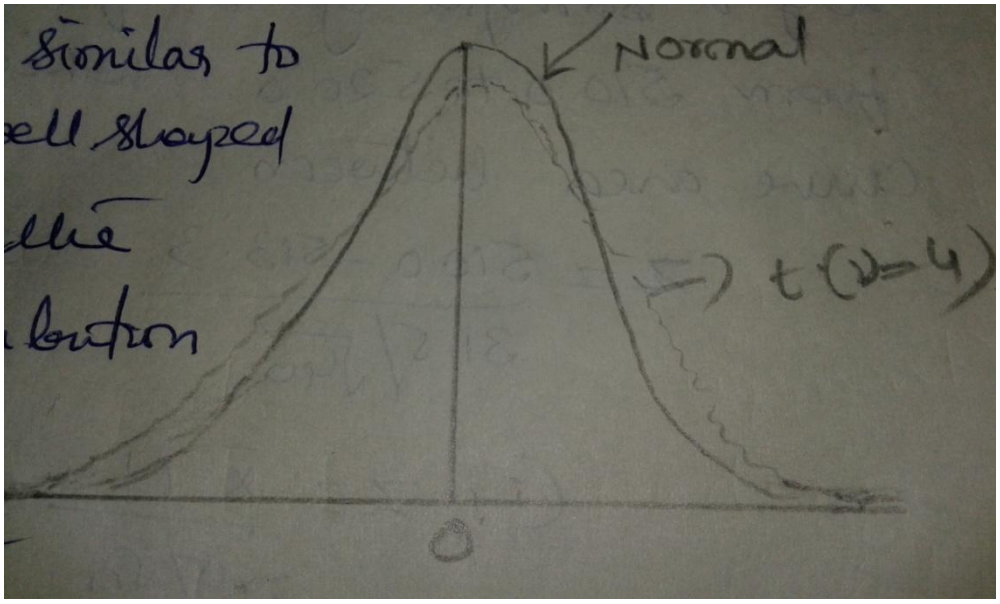
If n is large , this does not pose any problems even when σ is unknown, as it is reasonable in that case to substitute for it the sample standard deviation s. however, when it come stop the random variable whose values are given by $\frac{x - \mu}{\sigma / \sqrt{n}}$, very little is known about its exact sampling distribution ofr small values of n unless we make the assumption that the sample comes from a normal population under this assumption we have

If $\bar{x}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2 and

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} then$$

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

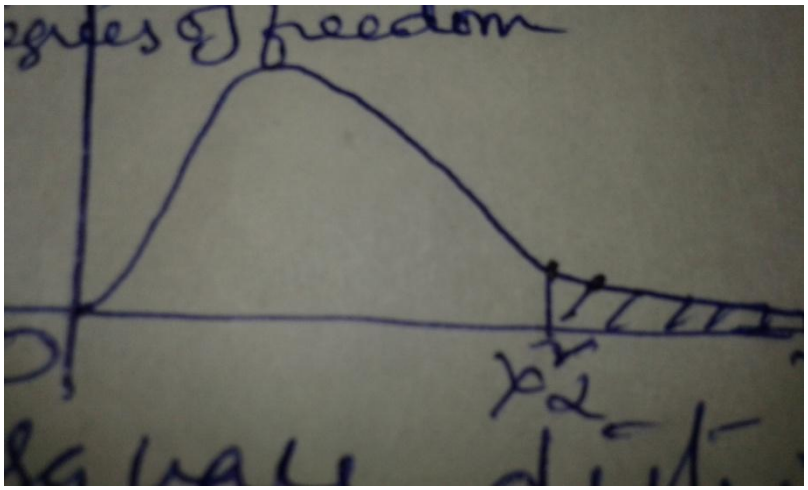
Is a random variable having the t distribution with the parameter $\gamma = n - 1.$ if can be seen from figure the overall shape of a t- distribution is similar to that of a normal distribution both are bell shaped and symmetrical about the mean. Take the standard normal distribution, the t- distribution has the mean 0, but its variance depends on the parameters $\gamma.$ called the number of degrees of freedom. The variance of the t- distribution exceeds 1, but it approaches 1 as $n \to \infty.$ In fact it can be shown that the t- distribution with degrees of freedom approaches the standard normal distribution as $\gamma \to \infty.$



Hence the standard normal distribution provides a good approximation to the t- distribution for samples of size 30Ω more.

Sampling Distribution of the variance:

Here we are concerned with the theoretical sampling distribution of the sample variance for random samples from normal populations. Since S^2 cannot be negative, we should suspect that this sampling distribution is not a normal wave in fact, it is related to the gamma distribution with $\alpha = \frac{d}{2}$ .and $\beta = 2$ is called the chi square distribution then we have if S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 then



$$\lambda^2 = \frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

Is a random variable having the chi-square distribution with the parameter $\gamma = n-1$.

Example: An optical firm purchases glass to be ground into lenses and it is known from part experience that the variance of the refractive index of this kind of glass is 1.26×10^{-4}. As it is important that the various pieces of glass have nearly the same index of refraction , the firm rejects such a shipment if the sample variances of 20 peaces selected at random exceeds 2.00×10^{-4}.Assuming that the sample values may be looked as a random sample from a normal population. What is the probability that a shipment will be rejected even though $\sigma^2 = 1.26x10^{-4}$?

Solution: Given n=20, $S^2=2.00 \times 10^{-4}$, $\sigma^2 = 1.26x10^{-4}$.

λ^2 or chi – square statistic i.e.,

$$\lambda^2 = \frac{(n-1)s^2}{\sigma^2} \sim v.d.f., v = n-1 d.f$$

$$\lambda^2 = \frac{(20-1)(2.00 \times 10^{-4})}{1.26 \times 10^{-4}} = 30.2$$

Since n = 20, V = 20.1 = 19 d.f at 5% level of significance $\lambda^2_{0.05} = 30\%$.

Thus , the probability that a good shipment will erroneously be rejected is less than 0.05. At 5% L.O.S v=19 d.f the $\lambda^2 - table.$ value is taken from $\lambda^2 - tables.$

2.3 Inferences on the mean and variance of a distribution

2.3.1 Hypothesis Testing.

In some problems we have to make the decisions whether a statement concerning a parameter is true or false, in order to estimate the value of a parameter we must test a hypothesis about a parameter. The hypothesis that is being tested is denoted as H. Here there are two possibilities whether H is true or false. If the hypothesis H is true and accepted or false and rejected, the decision is in either case correct. If the hypothesis H is true but rejected, here the rejection of H is an error. Also if hypothesis H is false but accepted then the acceptance of H is an error, then in the first case the error is called type I error and it is denoted by α , and in the second case the error is called type II error and it is denoted by β .

Hence the study of tests of significance enables us to divide on the bases of sample results if the deviation between the observed sample statistic and hypothetical parameter value orthe deviation between two independent sample statistics is significant or might be attributed to chance or the sampling fluctuations. We use the normal test of significance for large sample and for small samples the tests of significance based on t-test X Z-test and F-test.

Null Hypothesis:

Let us suppose that the bulbs manufactured under some standard manufacturing process have an average life of M hours and it is proposed to test a new procedure for manufacturing light bulb. Thus we have two populations of bulbs those manufactured by the standard process and new process in this problem the following three hypothesis may be set.

(i) New process is better than standard process.
(ii) New process is inferior to standard process.
(iii) There is no difference between the two processes.

The first two Statements appear to be biased since they reflect a preferential attitude to one or the other of the two processes. Hence the best course is to adopt the hypothesis of no difference, as stated in (iii). Thus null hypothesis is defined as a positive definite statement about the population parameter under consideration such a hypothesis is called hypothesis of no difference and usually denoted by H0. And

(i) $\mu_0 : \mu = \mu_0, (ii), H_0 : \sigma^2 = \sigma_0^2 etc.$

Alternative Hypothesis:
Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis usually denoted by H1.
For example, if we want to test the null hypothesis that the population has a specified mean μ_0 says, i.e., $H_0 : \mu = \mu_0$, then the hypothesis could be

(i) $H_1 : \mu \neq \mu_0 (i.e., \mu > \mu_0 or \mu < \mu_0)$
(ii) $H_1 : \mu > \mu_0$

(iii) $H_1 : \mu < \mu_0$

The alternative hypothesis in (i) is known as a two tasted alternative and the alternatives in (ii) and (iii) are known as right tailed and left tailed alternatives respectively.

Type 1 Error:

The probability of rejecting H0 when it is true is called as type I error. And it is denoted by α and written as

P(Reject H0 when it is true) = α

Also called as producers risk.

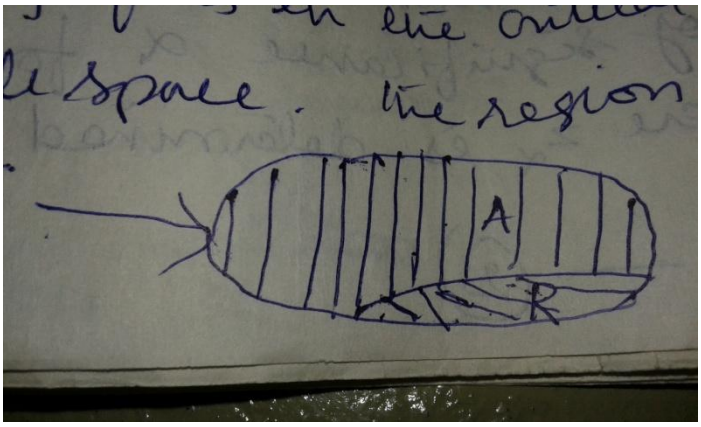
Type II Error:

The probability of accepting H_0 when it is wrong is called as type II error and it is denoted by β and written as

P (Accept H_0 when it is wrong) = β

Also called as consumers risk

Critical Region and Acceptance Region :



Suppose the sample values $x_1, x_2, ...x_n$ determine a point E on the n dimensional sample space S which would be the set of the various sample points corresponding to the all possible outcomes of the experiment the testing of statistical hypothesis is made on the basis of the division of this sample space into two mutually exclusion regions one region for acceptance (acceptance region) and another for the rejection critical region of H_0. The null hypothesis is rejected as soon as the sample point falls in the critical or in the rejection region of the sample space. The region of rejection is denoted by R or By C .critical region $R \leq C$ and acceptance region $A < S, S = A + R$ the null hypothesis is accepted as soon as the sample point falls in the acceptance region, which is denoted by A.

Level of significance.

The maximum value of type I error which we would be willing to risk is called level of significance of the test in practice 0.05 and 0.01 are the commonly accepted values of it simply means that on the average in 5 chances out of 100 we are likely to reject a correct H_0. Also with this we would be 95% confident of our decision in rejecting a null hypothesis.

Critical values or significant values.

The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value of significance value. It depends on

(i)The level of significance used and

(ii) The alternative hypothesis, whether it is two tailed or single tailed.

In case of large sample for example the standardized variable corresponding to the statistic t is

$$Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1).....(1)$$

Asymptotically as $n \to \infty$ the value of Z given (1) under the null hypothesis is known as test statistic the critical value of the test statistic at level of significance α for a two tailed left is given by Z_α, where Z_α is determined by the equation.

$$P(|Z| > Z_\alpha) = \alpha ----------(2)$$

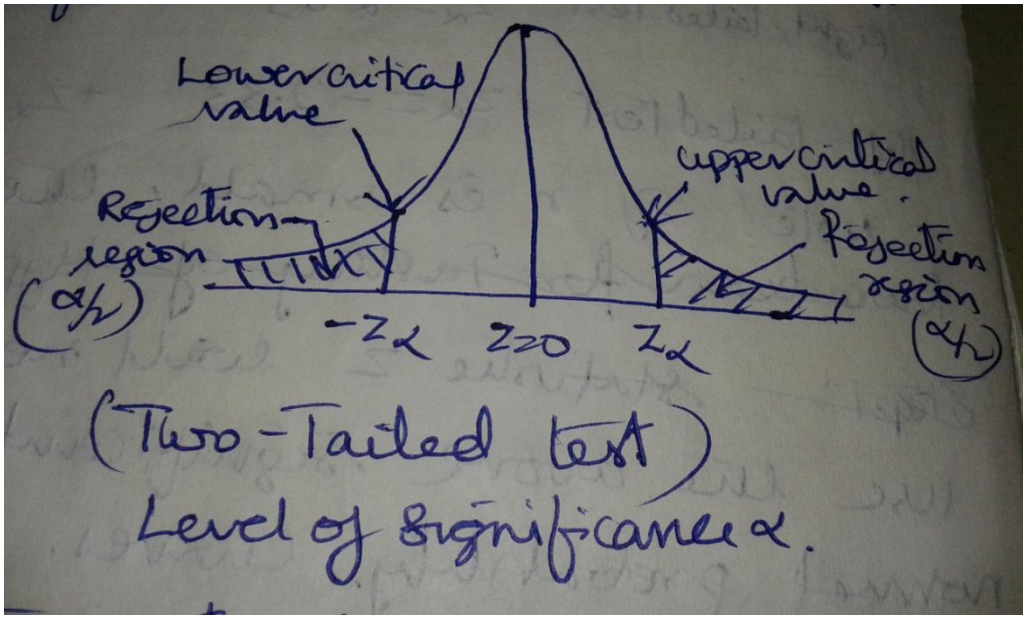
i.e Z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is symmetrical curve from 2 we get.

$$P(Z > Z_\alpha) + p(Z < -Z_\alpha) = \alpha$$

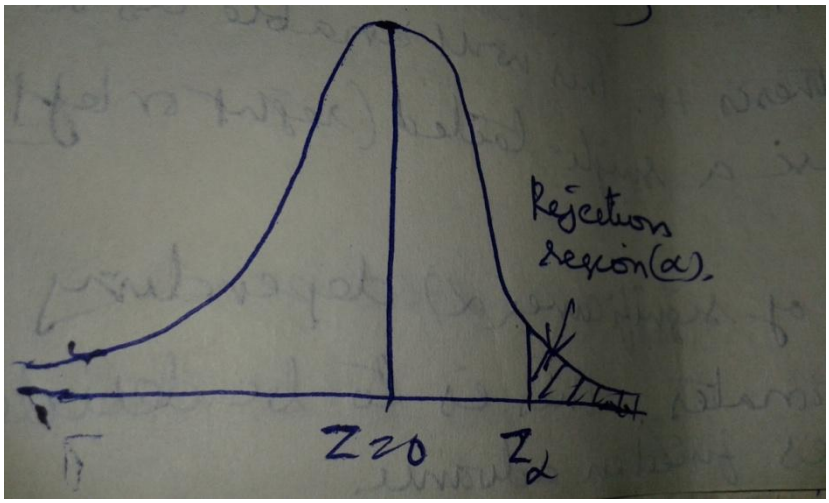
$$\Rightarrow p(Z > Z\alpha) + p(Z > Z_\alpha) = \alpha$$

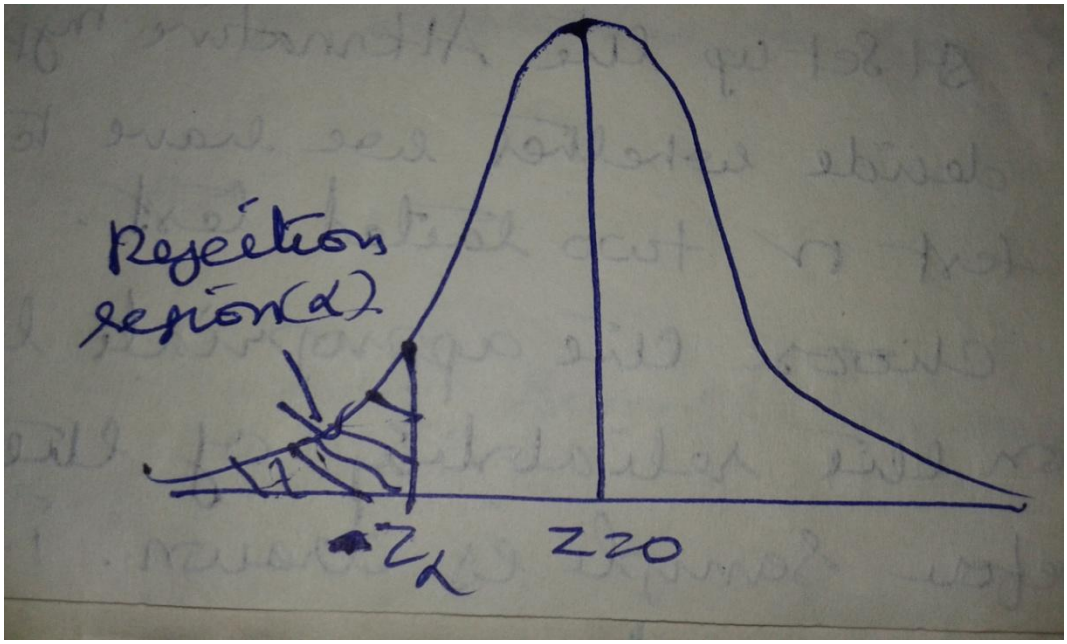
$$\Rightarrow 2p(Z > Z_\alpha) = \alpha/2$$

i.e the area of each tail is $\alpha/2$. thus Z_α is the value such that are to the right of Z_α is $\alpha/2$ and to the left of $-Z_\alpha$ is $\alpha/2$ as shown in the figure.



In case of single alternative, the article value Z_α is determined so that total area to the right of it (for right tailed test) is α and for left tailed test the total area to the left of $-Z_\alpha$ is α as shown in the figures. For right tail test $p(Z > Z_\alpha) = \alpha$, for left tail test $p(Z < -Z_\alpha) = \alpha ----(4)$





Thus the significance or critical value of Z for a single tailed test (left or right) at level of significance α is same as the critical value of Z for a two tailed test at level of significance '2 α .Here we give the critical values of Z at commonly used levels of significance for both two tailed and single tailed tests . These values have been obtained from equations 2 and 3 and 4 on using the normal probability tables.

Critical values (Z_α)	Level of significance (α) 1% 5% 10%
Two tailed test	$\|Z_\alpha\| = 2.58 \quad \|Z_\alpha\| = 1.96 \quad \|Z_\alpha\| = 1.645$
Right tailed test	$Z_\alpha = 2.33 \quad Z_\alpha = 1.645 \quad Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33 + Z = -1.645 \quad Z_\alpha = -1.28$

If n is small , then the sampling distribution of the test statistic Z will not be normal and in that case we can't use the above significant values which have been obtained from normal probability curves. In this case n is small usually less than 30 we use the significant values based on the t, X^2 ,F tests. The significant values of these tests for different values of α and γ degrees of freedom are given separately

Procedure for testing of hypothesis

Step 1: Set up the Null hypothesis H_0

Step 2: Set up the alternative hypothesis H_1. This will enable us to decide whether we have to use a single tailed (right or left) test n two tailed test.

Step 3: choose the appropriate level of significance α depending on the reliability of the estimates this is to be decoded before sample is drawn i.e. α is fixed in advance.

Step 4: Compute the test statistic

Under the null hypothesis.

Step 5: We compare the Z calculated value in step 4 with the significant value tabulated value Z_α at the given level of significance α .

If $|Z| < Z_\alpha$ i.e., if the calculated value of Z in modulus is less then Z_α we say it is not significant. Hence we accept the null hypothesis at α level of significance.

If $|Z| > Z_\alpha$ i.e, if the calculated value of Z in modules is greater than Z_α we say that it is significant and the null hypothesis is rejected a α. Level of significances.

2.3.3 Hypothesis and significance test on the mean:

Let $x_1, x_2, \ldots\ldots x_n$ is a random sample of size n being taken from a normal population with mean μ and variance σ^2 then the sample mean is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$ i.e., $\bar{x} \sim N(\mu, \sigma^2/n)$.

However this result olds i.e., $\bar{x} \sim N(\mu, \sigma^2/n)$. even for random sample taken from a non normal population provided the sample size n is large (central limit theorem). Then the test statistic to test the null hypothesis is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

For small samples when σ is unknown sample variance can be calculated by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The test statistic to test the null hypothesis $H_0 : \mu = \mu_0$ for single mean in case of small samples is given by

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} d.f$$

Which is a random variable having the t- distributed with (n-1) degrees of freedom. Comparing the computed t value and tabulated t- value we arrive at the following conclusions.

i.e., if $c_{alculated}$ value is less than $t_{abulated}$ value at a specified level of significance α based on two tailed or single tailed tests the null hypothesis is accepted.

If $c_{alculated}$ value is greater than $t_{tabulated}$ value at a specified level of significance α based on two tailed or single tailed tests the null hypothesis is rejected. And the value of t- tabulated value is taken from t- tables based on v=n-1 d.f and specified level of significance α

2.3.4 Hypothesis tests on the variance:

Consider the problem of testing the null hypothesis that a population variance equals a specified constant against a suitable one sided or two sided alternative that is we shall test the null hypothesis $\sigma^2 = \sigma_0^2$ against one of the alternatives $\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$ or $\sigma^2 \neq \sigma_0^2$. Tests like these are important whenever it is desired to control the uniformly of a product or an operation for example, suppose that a silicon disc or water is to becut into small squares or dice to be used in the manufacture of a Semiconductor device .since certain electrical characteristics of the finished device may depend on the thickness of the due it is important that all dice cut from a wafer have approximately the same thickness. Thus not only must the mean thickness of a wafer be kept within specifications, but also the variation in thickness from location to location on the wafer.

Based on the fact that for random samples from a normal population with the variance σ_0^2, here we test the null hypothesis that population variance equals a specified constant against a suitable one sided or two sided alternatives, with a specified level of significance α .Then the test statistic to test the null hypothesis H_0 is given by

$$\lambda^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim (n-1)d.f$$

Which is a random variable having the (λ^2) chi square distribution with (n-1) degrees of freedom. Based on (n-1) d.f and a specified level of significance α that chi square tabulated value is obtained from λ^2

Tables. Now on comparing the chi square calculated value and chi square tabulated value we come to the following conclusions.

If λ^2 calculated value < χ^2 tabulated at (n-1) d.f with a specified level of significance α we accept our null hypothesis H_0 and if χ^2 calculated value > χ^2 tabulated at (n-1) d.f with a specified level of significance α we reject our null hypothesis H_0

2.4 Inference on proportions:

2.4.1 Estimating proportions:

If x is the number of successes in n independent trials with constant probability P of success for each trial and if n trials satisfy the assumptions underlying the binomial distribution, then the mean and standard deviation of the number of success are given by Np and

$\sqrt{np(1-p)}$ where p is the probability of success.

i.e., E(x)=np, V(x)=np(1-p) (Where 1-p is the probability of failure

$$S.D(X) = \sigma = \sqrt{V(X)} = \sqrt{np(1-p)}$$

If we divide E(x) and S.D(x) (σ) by n

We get

$$E(\frac{X}{n}) = \frac{1}{n}E(X) = \frac{1}{n}.np = p$$

$$\frac{\sigma}{n} = \frac{\sqrt{np(1-p)}}{n}$$

Also as sample proportion 'p' gives an unbiased estimate of the population proportion p and if will p for probability of success and Q as probability of facture.

i.e., Q=1-p. p= sample proportion.

$$\therefore E(p) = E\left[\frac{x}{n}\right] = p$$

$$V(p) = V(\frac{x}{n}) = \frac{1}{n^2}V(x) = \frac{npQ}{n^2} = \frac{PQ}{n}$$

$$\therefore Stndarderrorofp = \sqrt{V(p)} = \sqrt{\frac{PQ}{n}}$$

Note : If the sample is taken from a finite population of size N then standard error of proportions is

S.E(P)= $\sqrt{\frac{(N-n)}{(n-1)}.\frac{PQ}{n}}$

To find the confidence interval for p

From the definition of binomial distribution we know that the probability of X success out of N independent trials with constant probability P of success for each trial is

$$p(x) = n_{c_x} - p^x.Q^{n-x}; x = 0,1,2,...n$$

And mean E(x)=np and V(X)=nPQ Where Q=1-p.

For large n, the binomial distribution tends to normal distribution. Hence for large n, X~n(np,Npq)

i.e., $Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{nPQ}} \sim N(0,1) or X - np$

Then confidence limits for p in terms of the observed values X and substituting x/n for p. we have.

$$\frac{x}{n} - Z_{\alpha/2}\sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}} < p < \frac{x}{n} + Z_{\alpha/2}\sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}$$

Is the confidence interval for p. if write $\frac{x}{n} = p$, the confidence interval for large sample for p is

$$p - Z_{\alpha/2}\sqrt{\frac{PQ}{n}} < 1 < p + Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

Maximum Error:

When we use a sample proportion to estimate the population proportion , we know that although we are using a method of estimation which has certain desirable properties , the chances are stem, virtually nonexistent, that the estimate will actually equal p. Hence, it would seem describable to accompany such a point estimate p with some statement as to how close reasonably expect the estimate to be. The error X/n-p is the difference between the estimator and the quantity it is supposed to estimate. In order to examine this error we make use of the fact that for large n.

The inequality $\left|\frac{x}{n} - p\right| \le Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$ will be satisfied i.e., the error will be at most $Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

Maximum error of estimate for the proportion p is $E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

Determination of sample size :

Considering maximum error of estimate for the proportion p is

$$E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

$$\sqrt{n} = Z_{\alpha/2}\sqrt{\frac{PQ}{E}}$$

Hence $n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 .(PQ)$

Sample size $n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 .(PQ)$ or $n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 p(1-p)$

Note : If P is not given , the above formula cannot be used if p is not given we can make use of the fact that p(1-p) is atmost ¼ .

Sample size n (when p is not given)=1/4 $n = \left(\frac{Z_{\alpha/2}}{E}\right)^2$

2.4.2 Testing hypothesis on a proportion:

Here we shall test the null hypothesis $H_0 = p = p_0$ against one of the alternatives $P < P_0, p > P_0, or p \neq p_0$ we use the Z- statistic . i.e., we shall consider only approximate

large sample list based on the normal approximate to the binomial distribution. than the test statistic for large sample test concerning p is

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$$

Which is a random variable having approximately the standard normal distribution. The critical values are same values that were used for test concerning one mean

If we compare Z calculated value and Z tabulated value for a specifically level of significance α ,we reject H_0 if Z cal>Ztab otherwise we accept H_0 if z cal< Z tab.

2.4.3 Estimating two proportions :

Suppose we want to compare two distinct populations with respect to the prevalence of certain attribute, say A among their members .let x1,X2 be the number of persons possessing the given attribute A in random samples of sizes n1,n2 taken from two populations respectively then sample proportions are given by

$$P_1 = \frac{X_1}{n_1}, p_2 = \frac{X_2}{n_2}$$

If P1 and P2 are population proportions then

$$E(p_1) = p_1, E(p_2) = p_2$$

$$v(p_1) = \frac{P_1Q_1}{n_1}.V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for large samples P1 and P2 are asymptotically normally distribution (P1-P2) is also normally distributed the standard variable corresponding to the difference (p1-p2) is given by

$$Z^1 = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0,1)$$

Under the null hypothesis

$H_0 : P_1 = P_2$ i.e. there is no significant difference between sample proportions.

$$E(p_1 - p_2) = E(p_1) - E(p_2) = p_1 - p_2 = 0$$

$$V(p_1 - p_2) = V(p_1) - V(p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} = PQ(\frac{1}{n_1} + \frac{1}{n_2})$$

$$S.E(p_1 - p_2) = \sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}$$

The confidence interval for difference of two proposition is given by

$$(p_1 - p_2) \pm Z_{\alpha/2}\sqrt{PQ(\frac{1}{n_1}+\frac{1}{n_2})}$$

Note: Suppose the population proportions P1 and P2 are given to be different i.e., $P_1 \neq P_2$ and we want to test whether the difference $P_1 - P_2$ Significant

And $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$

$$E(p) = \frac{1}{n_1 + n_2} E\left[n_1 p_1 + n_2 p_2\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 E(p_1) + n_2 E(p_2)\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 p_1 + n_2 p_2\right]$$

The estimate is unbiased.

2.4.4 Testing of hypothesis of two proportions:

When we compare the consumer response percentage favorable and percentage in favorable to two different products, when we decide whether the proportion of defectives of a given process remains constant from day to day, when we judge whether there is a difference in political persuasion among several nationality groups, and in many similar situations, we are interested in testing whether two binomial populations have the same parameter p. we are interested in testing the null hypothesis

$$H_0 : P_1 = P_2 = P$$

Against the alternative hypothesis that this population proportions are not equal. To perform a suitable large sample test of this hypothesis, we require independent random samples of size n_1, n_2, then, if the corresponding number of successes are x_1 and x_2. Under $H_0 : P_1 = P_2$ the test statistic for difference of proportions becomes.

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)$$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

And the estimate is unbiased, therefore E(p)=p which is the required test statistic. Company the Z calculated value and Z tabulated value we arrive at a conclusion that if $Z_{cal} < Z_{tab}$ at a specified level of significance we are going to accept H_0 at α level of significance.

Note: Suppose the population proportions P1 and P2 are given to be different i.e., $P_1 \neq P_2$ and we want to test whether the difference $P_1 - P_2$ Significant

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{S.E(p_1 - p_2)} = \frac{(p_1 - p_2) - (p_1 - p_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

Here the sample proportions are not given under the null hypothesis $H_0 : P_1 - P_2$, the differences in population.proportions is likely to be zero. Then the statistic becomes.

$$Z = \frac{|p_1 - p_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2})}$$

$$S.E(p_1 - p_2) = \sqrt{(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2})}$$

And the maximum error = $Z_{\alpha/2} \sqrt{(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2})}$

2.5 worked out example:

Example: the mean of random sample is an unbiased estimate of the mean of the population = 3,6,9,15,27,list all possible samples of size 3 that can be taken with outreplacement from the finite population.

(a) Calculate the mean of each of the samples listed in and assigning each sample a probability of 1/10. Verify that the mean of these $\bar{x}$ Is equal to 12, which is equal to the mean of the population θ.i.e., $E(\bar{x}) = \theta$.i.e. prove that $\bar{x}$ is an unbiased estimate of θ.

Solution : List of all possible samples of size 3 taken from 3,6,9,15,17 without replacement ($5_{c_3} = 10$, since there are samples) = (3,6,9), (3,6,15), (3,6,17),(6,9,15), (6,9,17), (3,9,15), (3,9,17), (9,15,17), (6,15,17), (3,15,17).

Mean of the population $\theta = \frac{3+6+9+15+27}{5} = 12$

Mean of samples =6,8,12,10,14,9,13,17,16,15

Probability assigned is 1/10 each.

$\bar{x}$	6	8	12	10	14	9	13	17	16	15
$p(\bar{x})$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

$$E(\bar{x}) = 6.\frac{1}{10} + 8.\frac{1}{10} + 12.\frac{1}{10} + 10.\frac{1}{10} + 14.\frac{1}{10} + 9.\frac{1}{10} + 13.\frac{1}{10} + 17.\frac{1}{10} + 16.\frac{1}{10} + 15.\frac{1}{10}$$

$$= \frac{1}{10} \times 120 = 12 = \theta$$

$$E(\bar{x}) = \theta$$

$\bar{x}$ is an unbiased estimate of θ.

Hence the mean of a random sample is an unbiased estimator of the mean of the population.

Example 2: To estimate the average time it takes to assembles a certain computer component, the industrial engineers at an electronics firm timed 40 technics in the performance of the task getting a mean of 12.73 minutes and a standard deviation of 2.06+ minutes.

A what can we say with 99% confidence about the maximum error if $\bar{x}$ =12.73 is used as a point estimate of the actual average time required to do the job.

Use the given data to construct a 99% confidence interval.

What confidence can be assert that the sample mean does not differ solution.

a. Maximum error = $Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

   Given n=sample size =40
   $\sigma =$ standard deviation of the population =2.06

$Z_{\alpha/2} = 2.575$

Maximum error E = $\frac{2.575 \times 2.06}{\sqrt{40}} = 0.842$

Confidence interval = $\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$\bar{x} = 12.73$ $E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 0.842$

Confidence interval = (12.73-0.842,12.73+0.842)

= (11.888, 13.572)

Hence we have to find (α) 100% to find this , first find maximum error =30 seconds =0.5 minutes =E, $\sigma = 2.06$

$$E = Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \Rightarrow Z_{\alpha/2} = \frac{E.\sqrt{n}}{\sigma} = \frac{0.5\times\sqrt{40}}{2.06}, n = 40$$

$$= 1.53$$

$$Z_{\alpha/2} = 1.53$$

The area when Z=1.53 from tabulates is 0.437

$$\frac{\alpha}{2} = 0.437 \Rightarrow 0.874 = \alpha$$

$$\therefore \alpha 100\% = 87.4$$

We are 87.4% confidence that the maximum error is 30 sec.

Example 3: Find 95% confidence limits for the mean of a normality distributed population from which the following sample was taken 15,17,10,18,16,9,7,11,13,14.

$$\bar{x} = \frac{15+17+10+18+16+9+7+11+13+14}{10} = 13$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \left[\begin{array}{l}(15-13)^2 + (17-13)^2 + (10-15)^2 + (18-13)^2 + (16-13)^2 + (9-13)^2 + (7-13)^2 + (11-13)^2 + \\ (13-13)^2 + (14-13)^2\end{array}\right]$$

$$S^2 = 13.3 = \frac{40}{3}$$

$$t_{\alpha/2} = 2.26$$

$$t_{\alpha/2}\sqrt{\frac{s^2}{n}} = 2.26\sqrt{\frac{40}{3}}.\frac{1}{10} = \frac{2.26\times 2}{\sqrt{3}} = 2.6$$

Confidence limits are $\bar{x} - t_{\alpha/2}\sqrt{\frac{s^2}{n}}, \bar{x} + t_{\alpha/2}\sqrt{\frac{S^2}{n}}$

=(13-2.6,13+2.6)

=(10.4,15.6)

Example 4: A random sample of size 16 values from a normal population showed a mean of 41.5 inches and the sum of the squares of deviations form mean is 135 sq. inches. Find the maximum error with 95% confidence

Solution : Maximum error = $t_{\alpha/2}\sqrt{\frac{s^2}{n}}$

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}, given that \sum_i (x_i - \bar{x})^2 = 135$$

$$n = 16, S^2 = \frac{135}{15} = 9, s = 3, t_{\alpha/2} = 2.131$$

$$E = t_{\alpha/2} . \frac{s}{\sqrt{n}} = 2.131 . \frac{3}{\sqrt{16}} = 1.598$$

Maximum error =1.598

Example 5: Measurements of the weights of a random sample of 200 ball bearings made by a certain machine during one week showed a mean of 0.824 and a standard deviation of 0.042. find 95% confidence limits for the mean weight of all the ball bearings.

Solution : confidence limits $(\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$

$\bar{\chi} =$

mean of the sample =0.824

$Z_{\alpha/2}$

, = Z value for 95% level (from tables)=1.96

$\sigma =$

standard deviation =0.042

N=size of the sample =200

$$Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \frac{1.96 \times 0.042}{\sqrt{200}} = 0.0059$$

$$(\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 0.824 - 0.0059 = 0.8181$$

$$\bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 0.824 + 0.0059 = 0.8299$$

$\therefore$

Confidence limits are (0.8181,0.8299)

Example 6: According to the norms established for a mechanical aptitude test, persons who are 18 years old have an average height of 73.2 with a standard deviation of 8.6. if 45

randomly selected persons of that age averaged 76.7. test the null hypothesis μ=73.2, against the alternative hypothesis is μ>73.2 at the 0.01 level of significance

Solution: Null hypothesis H_0:μ = 73.2

Alternative hypothesis H_1: μ> 73.2

Level of significance =99% or probability is 0.01

Test statistic : $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

μ = 73.2, $\bar{x}$ = mean of the sample =76.7

σ = standard deviation of population =8.6

N = sample size =45

$$z = \frac{76.7 - 73.2}{\frac{8.6}{\sqrt{45}}} = 2.73$$

Table value z_α = 2.33

Since Z calculated value 2.73 is > Z tabulated value 2.33 at 1% level of significance we are going to reject the null hypothesis. That is μ> 73.2, at the 1% level of significance.

Example 7: A random sample of six steel beams has a mean compressive strength of 58392 with a standard deviation of 648. Use this information at 5% level of significance test whether the true average compressive strength of the steel from which this sample came is 58,000

Null hypothesis $H_0 : \mu = 58{,}000$

Alternative hypothesis $H_1 : \mu \neq 58{,}000$

Level of significance $\alpha = 0.05$

Test statistic to list the null hypothesis $\frac{t - \bar{x} - \mu}{s/\sqrt{n}} \sim tn - 1d.f$

Given

$\bar{x}$ = sample mean =58392

S = Standard deviation = 648

$\mu = 58{,}000$ – mean of the population be tested

N= sample size =6

N<30, this is small sample with n-1=6-1=5 degrees of freedom.

Table value for $\alpha = 0.05$ and v=5, is 2.57= $t_{\alpha/2}$

$$t = \frac{58392 - 58{,}000}{\frac{648}{\sqrt{6}}} = 1.49$$

Since t calculated value 1.49 is less than 2.57 at 5% level of significance H_0 is accepted. Therefore mean of the population $\mu = 58{,}000$

Example 8: A random samples of 10 bags of pesticides are taken whose weights are 50,49,52,44,45,48,46,45,49,45, in kegs .test whether the average packing can be taken to be 50 kegs.

Null hypothesis $H_0 : \mu = 50$

Alternative hypothesis $H_1 : \mu \neq 50$

Level of significance $\alpha = 0.05$

Test statistic to list the null hypothesis $\frac{t - \bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1} d.f$

$$\bar{x} = \frac{50 + 49 + 51 + 44 + 45 + 48 + 46 + 45 + 49 + 45}{10} = \frac{472}{10} = 47.2$$

$$S^2 = \frac{\sum (x_i - \bar{x})}{n-1} = (50 - 47.2)^2 + (49 - 47.2)^2 + ... + (45 - 47.2)^2$$

x_i	50	49	51	44	45	48	46	45	49	46
$\bar{x}$	20.8	1.8	3.8	-3.2	-2.2	0.8	-1.2	-2.2	1.8	-1.2
$(x_i - \bar{x})^2$	7.84	3.24	14.44,10.24,4.84,0.64,1.44,					4.84,	3.24	1.44

$$\sum (x_i - \bar{x})^2 = 52.20$$

$$S^2 = \sum_i \frac{(x_i - \bar{x})^2}{n-1} = \frac{52.20}{9} = 5.8$$

$\bar{x}$ = mean of the sample =47.2

μ = mean of the population =50

N= sample size =10

$$t = \frac{47.2 - 50}{\sqrt{\frac{5.8}{10}}} = -3.6$$

$t_{\alpha/2} = 2.26$ table value at 9 d.f at $\alpha = 0.05$ level of significance

Since $|t| > t_{\alpha/2}$ i.e., t calculated value is greater than t tabulated value at 9 d.f and with 5% level of significance we are going to reject one H_0 null hypothesis . Hence the average packing cannot be taken as 50 kegs.

Example 9: In a random sample of 200 claims filed against an insurance company writing contusion insurance on case 84 exceeds 1200, construct a 95% confidence interval for the true proportion of claims filed against their insurances company that exceed 1200.

Solution :
$$p = \frac{X}{n} = \frac{84}{200} = 0.42, n = sizeofthesample = 200$$
$$Q = 1 - p = \frac{116}{200} = 0.58$$

Confidence interval for p is

$$P - Z_{\alpha/2}\sqrt{\frac{PQ}{n}} < p < p + Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

Where $Z_{\alpha/2} = 1.96$

$$0.42 - 1.96\sqrt{\frac{0.42 \times 0.58}{200}} < p < 0.42 + 1.96\sqrt{\frac{0.42 \times 0.58}{200}}$$
$$\Rightarrow 0.42 - 0.068 < p < 0.42 + 0.068$$
$$\Rightarrow 0.352 < p < 0.488$$

Example 10: In a random sample of 400 industrial accidents it was found that 231 were due to least partially to unsafe working conditions .what can we say with 95% confidence about the maximum error if we use the sample proportion to estimate the corresponding true proportion.

Solution: P=0.578, Q=0.422, n=400, $Z_{\alpha/2} = 1.96$ are given

Maximum error $E = Z_{\alpha/2}\sqrt{\frac{PQ}{n}} = 1.96\sqrt{\frac{0.578 \times 0.422}{400}}$

0.048

Maximum error E=0.048.

Example 11: In a study designed to investigate whether certain detonators used with explosives in coal mining meet the requirement that at least 90% will equate the explosive when charged. It is found that 174 of 200 detonators function properly. Test the null hypothesis p=0.9 against the alternative hypothesis p<0.9 at the 5% level of significance.

Solution :

Null hypothesis $H_0 : p = 0.9$

Alternative hypothesis $H_1 : p < 0.9$

Level of significance $\alpha = 0.05$

Test statistic to list the null hypothesis $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}} \sim N(0,1)$

Given p=0.87, Q=0.13, n=200

$$Z = \frac{0.87 - 0.9}{\sqrt{\dfrac{0.9 \times 0.13}{200}}} = \frac{0.0.3}{\sqrt{\dfrac{0.09}{200}}} = -1.41$$

$$Tablevalue Z_{\alpha} = -1.645$$

Since Z calculated value 1.41 is less than Z tabulated value 1.645 at 5% level of significance we accept null hypothesis H_0. i.e., there is no significant evidence.

To say that the given kind of detonator fails to meet the required standard.

Example 12: In a city A 20% of a random sample of 900 school boys had a certain slight physical defect. In another city B 18.5% of random sample of 1600 school boys had the same defect .1s the difference between the proportions significant at 0.05 level of significance.

Solution:

Null hypothesis $H_0 : p_1 = p_2$

Alternative hypothesis $H_1 : p_1 \neq p_2$

Level of significance $\alpha = 0.05$

Test statistic $Z = \dfrac{p_1 - p_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1)$

$$p_1 = \frac{X_1}{n_1}, p_2 = \frac{X_2}{n_2}, X_1 = 180, X_2 = 216, n_1 = 900, n_2 = 1600$$

$$p_1 = \frac{180}{900} = 0.2, p_2 = \frac{296}{1600} = 0.185$$

$$p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{476}{2500} = 0.19, Q = 1 - p = 0.81$$

$$Z = \frac{0.2 - 0.185}{\sqrt{0.19 \times 0.81\left(\dfrac{1}{900} + \dfrac{1}{1600}\right)}} = 0.9$$

Table value $Z_{\alpha/2} = 1.96$

Since Z calculated value 0.9 is less than Z tabulated value 1.96 at 5% level of significance we accept null hypothesis H_0. i.e., there is no significant differences between two proportions.

Example 13: The lapping process which is used to grind certain silicon wafers to the proper thickness is acceptable only if σ, the population standard deviation of the thickness of dice cut from the wafers, is at most 0.50mil. use the 0.05 level of significance to test the null hypothesis $\sigma = 0.50$ against the alternative hypothesis $\sigma > 0.50$, if the thickness of 15 dice cut from such wafers have a standard deviation of 0.64mil.

Solution:

Null hypothesis $H_0 : \sigma = 0.50$

Alternative hypothesis $H_1 : \sigma > 0.50$

Level of significance $\alpha = 0.05$

The table value of χ^2 distribution with 14 degrees of freedom at 5% level of significance is 23.685 from χ^2-tables. Then the test statistic to test the null hypothesis H_0 is

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} d.f$$

N=15, s=0.64, $\sigma_0^2 = (0.50)^2$

$$\chi^2 = \frac{(15-1)(0.64)^2}{(0.50)^2} = 22.94$$

Since χ^2 calculated value =22.96 does not exceed χ^2 tabulated value =23.685, at 5% level of significance the null hypothesis cannot be rejected even though the sample standard deviation exceeds 0.50, there is not sufficient evidence to conclude that the lapping process is unsatisfactory.

2.6 Exercise:

1. What is the maximum error one can expect to make with probability 0.9, when using the mean of a random sample of size n=64 to estimate the mean of a population with $\sigma^2 = 2.56$?

Ans: 0.329

2. ifs.d of the sample s=4, n=25. Construct 95% confidence interval if $\bar{x} = 35$

Ans: 33.352,36.648

3. it is desired to estimate the mean number of hours of continuous use until a certain computer will first repairs. If it can be assumed that $\sigma = 48 hours$ how large a sample be needed so that one will be able to assert with 90% confidence that the sample mean is off by atmost 10 hours.

Ans : n= 62

4. A sample of 900 members has a mean 3.4cms and S.D. 2.61cms. is this sample has been taken from 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 99% confidence limits of the true mean?

Ans Z=1.73, (3.23,3.57), (3.197,3.60.)

5. in a recent study 69 of 120 meteorites were observed to enter the earths atmosphere with a velouty of less than 26 miles per second. If we estimate the corresponding true proportion as p what canwe say with 955 confidence about the maximum error

Ans : E=0.088

6. Among 100 fish caught in a large laks, 18 were inedible due to the pollution of the environment. If we true proportion, with what confidence can we assert that the error of this estimate is at most 0.065.

Ans :91%

7. in a large consignment of oranges a random sample of 64 oranges revealed that 14 oranges were bad . if it season able to ensure that 20% of the oranges are bad.

Ans : 3.8

8. A die is thrown 256 times. An even digit turns up 150 times can we say that die is unbiased

Ans biased

9. A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustments before they could be shipped, while the same was true for 14 of 400 tractors produced on another assembly line at the 0.01 level of significance, does these support the claim that the second production that does superior work.

Ans :Z=2.37

10. When we take a sample from a infinite population what happens to the standard error of the mean if the sample size is

(i) increased from 50 to 200

Hint : S.E of the mean = $\frac{\sigma}{\sqrt{n}}$.

Sample size =n

Then consider $n_1 = 50, S.E = \frac{\sigma}{\sqrt{n_1}} = \frac{\sigma}{\sqrt{50}}, n_2 = 200, n_1$ is increased to 200

S.E= $\frac{\sigma}{\sqrt{n_2}} = \frac{\sigma}{\sqrt{200}} = \frac{\sigma}{\sqrt{4}\sqrt{50}} = \frac{\sigma}{2\sqrt{50}}$

If the sample size is increased from 50 to 200 the standard error will be divided by 2.

(ii) Decreased from 225 to 25.

11. An optical firm purchases glass to be ground into lenses and it is known from part experiences that the variances of the refractive index of this kind of glass is $1.26X10^{-4}$. As it is important that the various pieces of glass have nearly the same index of refraction the firm rejects such a shipment if the sample various of 20 pieces selected at random exceeds $2.00X10^{-4}$. Assuming that the sample values may be looked upon as a random sample from a normal population what is the probability that a shipment will be rejected even though $\sigma^2 1.26 \times 10^{-4}$

Ans: 30.2.

2.7 Summary

In this unit an attempt is made to explain the concepts of estimation and testing of hypothesis along with examples. The most important aspect is the sampling distributions and central limit theorem. A number of examples are worked out and a good number of exercises are also given.

2.8 Technical Terms:

Point estimation

Interval estimation

Inferences on one mean and variance

Inference on one proportion and two proportions

Central limit theorem.

UNIT III

ESTIMATION AND TEST OF HYPOTHESIS - II

Syllabus

Comparing two means point estimation independent samples, comparing the variance the F-distribution comparing means , variances equal, analysis of variance of one way classification fixed effects model, comparing variances, pair wise comparisons , randomized complete block design.

Objective :

This unit is prepared in such a way that after studying the material the student is expected to have a through comprehension of the above said syllabus which is the breath of any statistical investigation and analysis .also the student would be equipped with theoretical and practical aspects.

Structure of lesson:

3.1 Introduction:

The tests concerning the difference between two for example, if two methods of welding are being consider for use with railroad rails, we may take samples and decide which is better by comparing their mean strengths, also if a licensing examination is given to engineers who graduated from two different colleges, we may want to decide whether any observed differences between the means of the scores of the students from the two colleges is significant or whether it may be attributed to chance.

Experimentaldesign consists of three processes of planning experiments, analyzing the results and interpreting the results. The technique for making inferences is known as the analysis of variance. This powerful technique was developed by prof. R. A fisher for separation of the experimentally observed variance into a number of components traceable to specific sources. There are some underlying assumptions to every analysis and it is for the investigator to see that the experiment is performed in a manner so that these assumptions are satisfied .Thecomplete sequences of steps taken to ensure an objective analysis leading to valid references is called the design of experiment and is an important step in statistical analysis . the purpose of an experimental design is to obtain maximum information with the minimum cost and lab our.

3.2 Comparing two means and variances :

3.2.1 inferences on two means:

To test the significance for difference of means suppose we select two samples which are independent to test the significance. Let $\bar{x}_1$ be the mean of a random sample of size n1, drawn from a population of mean μ_1 and variance $\sigma_1^2 and\, \bar{x}_2$ be the mean of an independent random sample of size n2 drawn from another population with mean μ_2 and variance σ_2^2 since n1 and n2 are large we have.

$$\bar{x} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \bar{x_2} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

$\bar{x_1} - \bar{x}_2$ is the difference between the two independent normal variables and hence also a normal variable Z for $\bar{x_1} - \bar{x}_2$ we want to test the null hypothesis $\mu_1 - \mu_2 = \delta$ where δ is a specified constant.

If the distributions of two independent random variables have the means $\mu_1 and \mu_2$ and the variances $\sigma_1^2 and \sigma_2^2$, then the distribution of their sum (or differences has the mean $\mu_1 + \mu_2 or (\mu_2 - \mu_1)$ and the variances $\sigma_1^2 + \sigma_2^2$.

In testing the significance for the difference of two means, we shall consider the alternative hypothesis

$$\mu_1 - \mu_2 < \delta, \mu_1 - \mu_2 > \delta or \mu_1 - \mu_2 \neq \delta$$

To find the variances of the differences between the means of the two samples is

$$\sigma^2_{\bar{x_1}} = \frac{\sigma_1^2}{n_1}, \sigma^2_{\bar{x_2}} = \frac{\sigma_2^2}{n_2}$$

$$\therefore \sigma^2_{x_1 - x_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$z = \frac{\bar{x}_1 - \bar{x}_2 - E(x_1 - x_2)}{S.E(x_1 - x_2)}$$

Case (i) Under the null hypothesis $H_0 : \mu_1 = \mu_2$ there is no significant difference between the sample means

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2 = 0$$

$$\therefore z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Case (ii) If the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta, or, \mu_1 - \mu_2 < \delta, or, \mu_1 - \mu_2 > \delta, i.e, \mu_1 - \mu_2 \neq 0$, then the test statistic to test the H_0 is given by

$$\therefore z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

In the above said both the basis if calculated value is < tabulated Z critical value at a specified level of significance α we accept the null hypothesis

If calculated z value > tabulated Z critical value we reject our null hypothesis at α. Level of significance

Note : if $\sigma_1^2 = \sigma_2^2 = \sigma^2, and, H_0 : \mu_1 = \mu_2, then$

$$\therefore z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Critical region for testing $\mu_1 - \mu_2 = \delta$

Alternative Hypothesis	Reject null hypothesis
$\mu_1 - \mu_2 < \delta$	$Z < -Z_\alpha$
$\mu_1 - \mu_2 > \delta$	$Z > Z_\alpha$
$\mu_1 - \mu_2 \neq \delta$	$Z < -Z_{\alpha/2}$ $Z > Z_{\alpha/2}$

When $H_0 : \mu_1 - \mu_2, i.e,.\delta = 0$

Alternative hypothesis	Reject null hypothesis
$\mu_1 < \mu_2$	$Z \leq -Z_\alpha$
$\mu_1 > \mu_2$	$Z > Z_\alpha$
$\mu_1 \neq \mu_2$	$Z < -Z_{\alpha/2}$ $Z > Z_{\alpha/2}$

Note : the confidence limits for difference of means when n is large is given by

$$\bar{x}_1 - \bar{x}_2 \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Test of significance for the difference of means when (σ unknown)

Under the under null hypothesis H_0 the samples have been drawn form the normal populations with means μ_1, and, μ_2 and under the assumptions that the population variances are equal i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then the test statistic

$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sigma_{\bar{x}-\bar{y}}}$, where $\sigma^2_{\bar{x}-\bar{x}_{21}} = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ is a random variable having the t distribution with n_1+n_2-2 degrees to freedom. Where

$$S^2 = \frac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$\sum(x_1 - \bar{x}_1)^2 =$ Sum of the squared deviations from the mean for the first sample.

$\sum(x_2 - \bar{x}_2)^2 =$ Sum of the squared deviations from the mean for the second sample.

(n_1+n_2-2) degree of freedom (i.e., there are n_1-1 independent deviations from the mean in the first sample and similarly n_2-1 for 2nd sample and thus we have n_1+n_2-2 independent deviations from the mean to estimate the population variance.

$$\therefore t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)\sqrt{n_1 + n_2 - 2}}{\sqrt{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)}} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Or

$$t = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(n_1 - 1)^2 S_1^2 + (n_2 - 1)^2 s_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

Case (ii) when the null hypothesis is $\mu_1 = \mu_2$, or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$ then

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} d.f$$

Where $S^2 = \frac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{(n_1 + n_2 - 2)}$

Then small sample confidence interval for $\delta = \mu_1 - \mu_2$ is 100 (1-α)% confidence interval.

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{\sum(x_1-\bar{x}_1)+\sum(x_2-\bar{x}_2)^2}{(n_1+n_2-2)}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Where $V=n_1+n_2-2$ d.f

The paired sample t-test:

Not two samples but a pair of values before and after test will be given , take the difference of the values. Here the variables are not independent we can consider di-difference between the values as one variable.

Let us now consider the case when

(i) The sample sizes are equal i.e., n1=n2=n say and

(ii) The two samples are not independent but the sample observations are packed together, i.e., the pair of observations (x_i,y_i) (i=1,2,3,…n) corresponds to the same i^{th} sample unit the problem is to test if the sample means differ significantly or not

For example, suppose we want to test the efficacy of a particular drug, say, for ininducing sleep. Let x_i and y_i(i=1,2,3,…n) be the readings, in hours of sleep, on the i^{th} individual before and after the dry is seven respectively. Here instead of applying the difference of the means test we apply the paired t- test. Here we consider the increments, $d_i=x_i-y_i$ (i=1,2,3,….n) under the null hypothesis H_0: that increments are due to fluctuations of sampling , i.e., the drug is not responsible for these increments , the statistic to test the hypothesis is

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1} d.f$$

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, and, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i-\bar{d})^2$$

3.2.2 inferences on Two Variances

Suppose we want to test

(i) Whether two independent samples x_i(i=1,2,3,……n_2) have been drawn from the normal populations with the same variance σ^2 says or

(ii) Whether the two independent estimates of the population variance are homogenous or not.

Under the null hypothesis H_0 that $\sigma_x^2 = \sigma_y^2 = \sigma^2$ the population variances are equal or (ii) two independent estimates of the population variance are homogeneous , the statistic F is given by

$$F = \frac{S^2{}_x}{S_y^2}$$

Where $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$, $and, S_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \bar{y})^2$

Are unbiased estimates of the common population variance σ^2 obtained from two independent samples and follows a F-distribution with (V_1,V_2) d.f ,where $V_1=n_1-1$ and $v_2=n_2-1$.

3.3 Analysis of variance:

The analysis of variances is a powerful statistical tool for tests of significance. The equate procedure only for testing the significance of the difference between two sample means is the test of significance based on t- distribution suppose when we have to consider there or more samples to be considered at a time in a particular situation an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population i.e., they have the same mean, for example, five fertilizers are applied to four plots each of wheat and yield of wheat on each of the plot is given and we may be interested in finding out whether the effect of these fertilizers on the yields is significantly different or in otherwords whether the samples have come from the same normal population the answer to this problem is provided by the technique of analysis of variance. The basic purpose of this analysis of variance is to test the homogeneity of several means.

The term analysis of variances was introduced by prof. R.A.Fisher in 1920's to deal with problem in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as (i) assignable causes (ii) chance causes the variation due to chance causes is beyond the control of human hand and cannot be traced separately the variation due to assumable causes can be detected and measured.

Definition: According to prof.R.A.Fisher, analysis of variance is the "separation of variances ascribable to one group of causes from the variances ascribable to other group".by this technique the total variation in the sample data is expressed as the sum of its non negative components where each of these components is a measure of the variation due to some specific independent source or factor or cause.

For the validity of the F-test in ANOVA, the following assumptions are made.

(i) The observations are independent
(ii) Parent population from which observations are taken is normal.
(iii) Various treatment and environmental effects are additive in nature.

3.3.2 Analysis of variance of one way classification fixed effects model:

A fixed effect model is also known as analysis of variances model in this model, the investigator is concerned to draw inferences about t treatments involved in the experiment .if we take the model for a design in which k treatments are randomly assigned to n homogeneous units where i^{th} treatment is replicated $r_i$times such that $\sum_{i=1}^{k} r_i = n.$ then the statistical model with usual notations is

$x_{ij} = \mu + \alpha_i + E_j$, where (i=1,2,……x;j=1,2,….r_i) in case, the main interest lies only in estimating the effect of k treatments included in the experiment the above model is a

fixed effect model. For this model the assumption is that $\sum_{i=1}^{k} \alpha_i = 0$ for instance, the investigators study is confined to the effect of k.fertilizers or fertilizer doses, he has to choose a forced effect model. Therefore model(1) is suited to completely randomized design.

Therefore analysis of variance (ANOVA) utilizes F-test each component variances is tested against error variances and conclusions are drawn in the same way as evedo in F-test for equality of two variances.

ANOVA one way classification:

Let us suppose that N observations x_{ij}(i=1,2,…..k;j=1,2,…….n_i) of a random variable x are grouped , on some bases, into k classes of sizes n_1,n_2,…….n_k respectively $\left(N = \sum_{i=1}^{k} n_i \right)$ is exhibited below.

Mean	Total
X_{11} x_{12}…….$x_1 n_1 \bar{x}_1$	T_1
X_{21} x_{22}…….$x_2 n_2 \bar{x}_2$	T_2
. .	.
. .	.
. .	
X_{i1} x_{i2}……$x_i n_i \bar{x}_i$	T_i
. . . .	.
. . . .	.
. . . .	.
X_{k1} x_{k2}……$x_k n_k \bar{x}_k$	T_k

The total variation in the observations x_{ij} can be into the following two components due to different bases of classification, commonly known as treatments

The variation within the classes, i.e, the inherent variation of the random variable within the observations of a class.

The first type of variation is due to assignable causes which can be detected and controlled by human and the second type variation is due to chance causes which are beyond the control of human hand. The main object of analysis of variance technique is to examine if there is significant differences between the class means in view of the inherent variability within the separate classes.

The procedure to test the significance using ANOVA one way classification or to analysis the data is given in following steps.

Step 1: Set up the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$

Step 2: specify the level of significance α

Step 3: specify the total no. of observations N and find the grand total

$$G = \sum_i \sum_j x_{ij}$$

Step 4: Find the R.S.S = $\sum_i \sum_j x_{ij}^2$

Step 5: correction Factor (c.F) $= \frac{G^2}{N}$

Step 6: Calculate T.S.S = R.S.S – C.F (T.S.S = total sum of squares)

Step 7: Find treatment sum of squares $= \sum_{i=1}^{k} \left(\frac{T_i^2}{n_i} \right) - c.f = s_t^2$

Step 8: Then error sum of squares = total .s.s – treatment S.S= s_t^2 .

Step 9: Draw the ANOVA table for one way classified (C.R.D).

ANOVA TABLE:

Sources of variation	Sum of squares	d.f	Mean sum of squares	Variance ratio
Treatment	S_t^2	k-1	$S_t^2 = \frac{S_t^2}{(k-1)}$	$\frac{S_t^2}{S_t^2} = F(k-1, N-k)$
Error	S_E^2	N-k	$S_E^2 = \frac{S_E^2}{(N-K)}$	
Total	S_T^2	N-1		

At a specified level of significance 'α'(usually 5% or 1%) with (k-1,N-k) degrees of freedom F-tabulated value is obtained from F-tables.

Step 10: Conclusions are to be drawn based on the above analysis of the table.

If F calculated value is less than F-tabulated value with F (k-1,N-K)d.f at a specified level of significance accept the null hypothesis H_0.

If F calculated value is greater than $F_{\text{-tablulated}}$ value with (K-1,N-K)d.f reject the null hypothesis H_0 at $\alpha\%$ level of significance.

Note : 1 Degrees of Freedom (df) : The number of independent varieties which make up the statistic (from example t, χ^2, F) is known as the degrees of freedom (d.f) usually denoted by V. that is the number of degrees of freedom , in general, is the total number of observations less the number of independent constraints imposed on the observations. For example, if k is the number of independent constraints in a set of data of n observations then v=(n-k).

Critical difference (C.D):

If the treatments show significant effect then we would be interested to find out which pairs of treatments differ significantly. for this, instead of calculating students t for different pairs of treatment means , we calculate the least significant difference at the given level of significance this least difference is known as the critical difference (C.D) and C.D at $\alpha\%$ level of significance is given by

C.D= S.E of difference between two treatment means x $t_{\alpha\%}$ for error d.f

We have $\text{var}(\bar{x}_i - \bar{x}_j) = \frac{\sigma_e^2}{r_i} + \frac{\sigma_e^2}{n_j} = \sigma_e^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$

$$S.E(\bar{x}_i - \bar{x}_j) = \sigma_e\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$$

Hence C.D for $S.E(\bar{x}_i - \bar{x}_j) = t_{x\%}$ for error d.fXSE$\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$ Since SE^2 provides an in biased estimate of σ_e^2. If each treatment is replicated n times i.e, n_i=n, $\forall, i = 1,2,.....k,$ then C.D for difference of means = $t_{x\%}$ for error d.f $XSE\sqrt{\frac{2}{n}}$.

3.3.3 Randomised complete block design or (RBD):

In common parlance, randomized complete block design is spoken as randomized block design (R.B.D) the word complete is implicit. In field experimentation if the whole of the experimental area is not homogeneous and the fertility gradient is only one direction then a single method of controlling the variability of the experimental material consists in stratifying or grouping the whole area into relatively homogenous striate or sub groups or blocks perpendicular to the direction of fertility gradient now if the treatments are applied at random to relatively homogeneous units within each striate sub groups or block and perpendicular to the direction of fertility gradient. Now if the treatments are applied at random to relatively homogeneous units within each striate or block and replicated overall the blocks, the design is a randomized block design this design has been shown to be more efficient and accurate than complete randomized design for most types of experimental work. In R.B.D no restrictions are placed on the number of treatments or the number of replicates. And statistical analysis is single and raped R.B.D is not suitable for large number of treatment or for cases in which complete block contains considerable variability.

Analysis of R.B.D:

In a R.B.D a single observation is made on each of the experimental units, then its analysis is analogous to ANOVA for a two way classified data and its mean model becomes

$$x_{ij} = \mu + T_j + b_j + E_{ij}, (i = 1,2,...t,\ j = 1,2,....r)$$

Where X_{ij} is the response or the yield of experimental from i^{th} treatment and j^{th} block. μ is the general mean effect , is the effect due to i^{th} treatment, b_j is the effect due to j^{th}block or

replicate and E_{ij} is the error effect due to random component assumed to be independently normally distributed with mean zero and variances σ_e^2, E_{ij}i.d ~N(0, σ_e^2)

Let the suffix I refer to treatment and j refer to the N=r X t furnish the data for the comparison of the values in the following Z x r two way table.

	mean	Total
X_{11} x_{12}…….x_{ij} … x_{1r} $\bar{x}_1$		T_1
X_{21} x_{22}…….x_{2j}..x_{2r} $\bar{x}_2$		T_2
. .		.
. .		.
. .		
X_{i1} x_{i2}……$x_{ij...}x_{ir}$ $\bar{x}_i$		T_i
. . .	.	.
. . .	.	.
. . .	.	.
		T_k
X_{t1} x_{t2}……$x_{tj.....}x_{tr}$ $\bar{x}_t$		
		↓
$\bar{x}_{.1}$ $\bar{x}_{.2}$ …. $\bar{x}_{.j}$ … $\bar{x}_{.r}$	$\bar{x}$	G
$T_{.1}$ $T_{.2}$..$T_{.j}$…$T_{.r}$ →		

The procedure to test the significance using R.B.D (Analogous to ANOVA two way classification) is given in the following steps

Step 1: Set up the null hypothesis $H_t : J_1 = J_2 = ... = J_t$
$H_b : b_1 = b_2 = ... = b_r$

Step 2: specify the level of significance α.

Step 3: specify the total no. of observations N and find the grand total $G = \sum_i \sum_j x_{ij}$

Step 4: Find the R.S.S = $\sum_i \sum_j x_{ij}{}^2$

Step 5: correction Factor (C.F) $= \frac{G^2}{N}$

Step 6: Then calculate T.S.S = R.S.S –C.F (T.S.S =total of squares)

Step 7: Find the treatment sum of squares or sum of squares due to treatment. (S.S.T) $= \frac{1}{r}\sum_{i=1}^{r} T_i^2 - C.F = S_T^2$

Step 8: Find the block sum of squares or sum of squares due to blocks

$(S.S.B) = \frac{1}{t}\sum_{j=1}^{t} T_j^2 - C.F = S_B^2$

Step 9: Then error sum of squares = total sum of squares – sum of squares due to treatments – sum of squares due to blocks = S_E^2.

Step 10: Draw the ANOVA table for randomized block design (R.B.D).

ANOVA TABLE

Source of variation	Sum of squares	d.f	Mean sum of squares	Variance ratio
Treatment	S_T^2	t-1	$S_T^2 = \frac{S_T^2}{(t-1)}$	$E_T = \frac{S_T^2}{S_E^2} \sim F((t-1),(r-1),(t-1))$
Block	S_B^2	r-1	$S_B^2 = \frac{S_B^2}{(r-1)}$	$F_B = \frac{S_B^2}{S_E^2} \sim F(r-1,(r-1)(t-1)$
Error	S_E^2	(t-1)(r-1)	$S_E^2 = \frac{S_E^2}{(t-1)(r-1)}$	
Total		Rt-1		

At a specified level of significance α usually 5% or 1% with (t-1,(r-1)(t-1),(r-1,(r-1)(t-1) degrees of freedom F. tabulated value is obtained from F tables.

Step 10: Conclusions are to be drawn based on the above ANOVA table.

If F- calculated value is less than F –tabulated value with F(t-1,(r-1)(t-1) d.f at a specified level of significance α we accept our H_0. That is ,their is no significant difference due to treatments. Or if F- calculated value is greater than F- tabulated value with F(r-1,(r-1)(t-1)) d.f reject the null hypothesis H_0 at α% level of significance. That is there is significant differences due to treatments. Similarly we can draw the conclusions for the blocks

Note 1: Experiment : An experiment is a device or a means of getting an answer to the problem under consideration .

2. Treatments: Various objects of comparison in a comparative experiment comparative experiments are designed to compare the effect of two or more objects on some population characteristic are termed as treatments for example. In field experimentation different fertilizers or different varieties of crop or different methods of cultivation are the treatments.

3. Experimental unit: the smallest division of the experimental material to which we apply the treatments and on which we make observations on the variable under study is termed as experimental unit for example in the field experiments the plot of land is the experimental unit.

4. Blocks : In agricultural experiments most of the times we divide the whole experimental unit relatively homogenous sub groups or strata. These strata which are more uniform among themselves than the field as a whole, are known as blocks.

5. yield : The measurement of the variable under study on different experimental units (eg: plots in field experiments) are termed as yields.

6. Replication : Replication means the execution of an experiment more than once. In other words the repetition of treatments under investigation is known as replication

3.4 Worked out examples:

Example 1: An investigation of two kinds of photocopying equipment showed that 71 failures of the first kind of equipment took on the average 83.2 minutes to repair with a standard deviation of 19.3 minutes, while 75 failures of the second kind equipment took on the average 90.8 minutes to repair with a standard deviation of 21.4 minutes. Test the null hypothesis $\mu_1 = \mu_2 \neq 0$, the hypothesis that on the average it take an equal amount of time to repair either kind of equipment against the alternative hypothesis $\mu_1 = \mu_2 \neq 0$ at the level of significance $\alpha = 0.05$

Solution : Set up the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ alternative hypothesis

$H_1 : \mu_1 - \mu_2 \neq 0$, level of significance $\alpha = 0.05$

Test statistic to test the null hypothesis H_0 is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Given

$\bar{x}_1$ = mean of the first sample =83.2 minutes

$\bar{x}_2$ = mean of the second sample = 90.8 minutes.

σ_1^2 = variance of the first sample = $(19.3)^2$ sq. minutes

σ_2^2 = variances of the second sample = $(21.4)^2$sq.minutes.

n_1 = size of the first sample = 71

n_2 = size of the second sample = 75.

Based on $\alpha = 0.05$ level of significance the Z- tbale value is $Z_{\alpha/2} = 1.96$ since this is a two tailed test.

$$Z = \frac{83.2 - 90.8}{\sqrt{\frac{372.5}{71} + \frac{457.96}{75}}} = \frac{-7.6}{\sqrt{11.36}} = -2.2$$

$i.e., |z| = 2.2$

Conclusion : Since $|z|$ calculated value is 2.2 greater than Z-tabulated value 1.96 for two treaded test at 5% level of significances null hypothesis h0 is rejected therefore if does not take equal amount of time to repair either kind of equipment.

Example 2: Suppose that we want to investigate whether on the average man earn more than 20 perweek more than women in a certain industry. If sample data shown that 60 men earn on the average $\bar{x}_1 = 292.50$ per week with a standard deviation of 15.6 , while 60 women earn or average $\bar{x}_2 = 266.10$ per week with a standard deviation of 18.20, what can we conclude at the 0.01 level of significance.

Solution : set up the null hypothesis $H_0 = \mu_1 - \delta = \mu_2$

Alternative hypothesis $H_1 = \mu_1 - \delta > \mu_2$

Level of significance $\alpha = 0.01$

From H_1 it is an right tailed test and Z-table or critical value $Z_\alpha = 2.33$

The list statistic to test the null hypothesis H_0 is

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}} \sim N(0,1)$$

$n_1 = n_2 = n.$

Hence are given $\bar{x}_1 = 292.5 =$ mean earn of men

$\bar{x}_2 = 266.10 =$ mean earn of women.

$\sigma_1 = S.D$ ofearn of men 15.6

$\sigma_2 = S.D$ ofearn of women = 18.2

$\delta = 20$

$n_1 =$ sample size of first sample =60.i.e, no. of men.

$n_2 =$ sample size of second sample =60, i.e., no. of women

$n_1 = n_2 = n$

$$Z = \frac{292.5 - 266.1 - 20}{\sqrt{\frac{1}{60}(243.4 + 331.2)}} = \frac{6.4}{3.1} = 2.07$$

Conclusion : Since Z-calculated value 2.07 is less than $z_{\text{-tabulated}}$ value 2.33 hence we accept our null hypothesis H_0 : $\mu_1 - \delta = \mu_2$. At 1% level of significance.

Example3: The mean yield of two sets of plots and this variability are given below .Examine whether difference in the mean yields of the two acts of plots is significant.

Solution : Null hypothesis

	Set of 40 plots	Set of 60 plots
Mean yield per plot	1258	1243

S.D per plot	34	28

Conclusion: since Z-calculated value 2.3 is greater than t- tabulated value 1-96 at 5% level of significance we reject the null hypothesis H_0. i.e., there is significant difference between the mean of the two samples.

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Level of significance $\alpha = 0.05$

From H_1 since the test is a two tailed test the Z – critical value at 5% level of significance is $Z_{\alpha/2} = 1.96$

Then the test statistic to test the null hypothesis H_0 is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Given $\bar{x}_1 =$ mean of the first sample =1258

$\bar{x}_2 =$ mean of the second sample=1243

$\sigma_1^2 =$ variance of the first sample = $(34)^2$=1156

$\sigma_2^2 =$ variance of the second sample = 784

n_1= size of the first sample =40

n_2= size of the second sample = 60

$$Z = \frac{1258 - 1243}{\sqrt{\frac{1156}{40} + \frac{784}{60}}} = 2.3$$

Conclusion: Since Z- calculated value 2.3 is greater than Z table value 1.96 at 5% level of significance we reject the null hypothesis H0. i.e., there is significant difference between the means of the two samples.

Example 4: The standard deviations of two samples are 8 and 12. Sample sizes are 200 and 100 find the standard error of the difference between the means and also find the confidence interval at 5% level of significance .means of the sample are 60,50.

Solution : The $S.E(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$\sigma_1^2 =$ variances of the first sample =8^2=64.

$\sigma_2^2 =$ variance of the second sample =12^2=144.

n_1= size of the first sample = 200

n_2= size of the second sample = 100

$$S.E(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{64}{200} + \frac{144}{100}} = 1.33$$

Confidence interval or limits for differences of two means is given by

$$\bar{x}_1 - \bar{x}_2 \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

=[(60-50)-(1.96(1.33), (60-50)+1.96(1.33)]

= (7.4,12.6)

Example 5: Measuring specimen of nylon yarn taken from two spinning machines.it was found that mean denier of 9.67 with a standard deviation of 1.81 while 10 specimens from the second machines has a mean denier of 7.43 with a standard deviation of 1.48 assuming that the populations sampled are normal and have the same variance. Test the null hypothesis $\mu_1 - \mu_2 = 1.5$ against the alternative hypothesis $\mu_1 - \mu_2 > 1.5$ at 5% level of significance.

Solution: Set up null hypothesis $H_0 : \mu_1 - \mu_2 = 1.5$

Alternative hypothesis $H_1 : \mu_1 - \mu_2 > 1.5$

Level of significance $\alpha = 0.05$

From H_1 since it is a right tail test and at 5% level of significance critical value of t is $t_{\alpha/2} = 2.12$.

Given $\bar{x}_1 =$ mean of the first sample =9.67

$\bar{x}_2 =$ mean of the second sample =7.43

$\delta = 1.5$

$n_1 =$ first sample of size n_1=8

Second sample of size n_2=10

Variance of first sample $s_1^2 = (1.81)^2 = 3.276$

Variance of second sample $s_2^2 = (1.48)^2 = 2.19$

$$\therefore s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} = \frac{3.276 \times 7 + 2.19 \times 9}{8 + 10 - 2}$$

$$\because s^2 = \frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)}{n_1 + n_2 - 2}$$

also

$$\sum (x_1 - \bar{x}_1)^2 = s_1^2 (n_1 - 1)$$

$$\sum (x_2 - \bar{x}_2)^2 = s_2^2 (n_2 - 1)$$

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t_{n1+n2-2} d.f$$

$$= \frac{9.67 - 7.43 - 1.5}{2.665\sqrt{\frac{1}{8} + \frac{1}{10}}} = \frac{0.74 \times 4 \times \sqrt{80}}{6.53\sqrt{18}} = 0.96.$$

Conclusion: Since $t_{calculated}$ value 0.96 is less than $t_{tabulated}$ value 2.12 for a right tail test at 5% level of significance we accept our null hypothesis H_0. Therefore $\mu_1 - \mu_2 = 1.5$

Example 6: Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results.

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity (5 percent values of t for 11 degrees of freedom =2.20)

Solution: Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_0 : \mu_1 \neq \mu_2$

Level of significance $\alpha = 0.05$

At 5% level of significance for 11 degrees of freedom critical value of t for a two tailed test is given by $t_{\alpha/2} = 2.20$

Test statistic to test the null hypothesis is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n1+n2-2} d.f$$

$n_1 = sizeofthefirstsample = 7$

$n_2 = sizeofthe\,\mathrm{sec}ondsample = 6$

$$\because s = \frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)}{n_1 + n_2 - 2}$$

X1	X2	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
28	29	-3.3	10.89	0.83	0.689
30	30	-1.3	1.69	1.83	3.35
32	30	0.7	0.49	1.83	3.35
33	24	1.7	2.89	-4.17	17.39
33	27	1.7	2.89	-1.17	1.145
29	29	-2.3	5.29	0.83	0.689
34		2.7	7.29		
219	169		31.43		26.613

$$\bar{x}_1 = \frac{219}{7} = 31.3$$

$$\bar{x}_2 = \frac{169}{6} = 28.17$$

$$\because s^2 = \frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)}{n_1 + n_2 - 2}$$

$$= \frac{31.43 + 26.613}{11} = \frac{58.043}{11}$$

$$= 5.276$$

$$\therefore t = \frac{31.3 - 28.17}{2.3\sqrt{\left(\frac{1}{6} + \frac{1}{7}\right)}} = 2.5$$

Conclusion: since t calculated value 2.5 is greater then t calculated 2.2,at 5% level of significance we reject our null hypothesis.

Hence both the horses do not have the same running capacity.

Example 7: In a study of the effective ness of physical exercise in weight reduction a group on 16 person engaged in a presented program of physical exercise for month showed the following result.

Weight before	209	178	169	212	180	192	159	180	170	153
Weight after:	196	171	160	207	177	190	158	180	164	152
(exercise)		183	165	201	179	243	144			
		179	162	199	173	231	140			

Use 0.01 level of significance to test whether the prescribed program of exercise is effective.

Solution: set up the null hypothesis $H_0 : \mu = 0$

(where μ is the mean of the population of differences sampled)

Alternative hypothesis $H_1 : \mu > 0$

Level of significance $\alpha = 0.01$

At 1% level of significance with v=n-1=16-1=15.

Degrees of freedom critical value of t is $t_\alpha = 2.6$

The test statistic to test the null hypothesis H_0 is

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1} d.f$$

$$where \bar{d} = \frac{\sum d_i}{n}, d_i = x_i - y_i$$

Where d is the mean of the differences

S is the standard deviation of the differences (i.e, difference between before and after exercise)

d_i=13,7,9,5,3,2,1,0,6,1,4,3,2,6,12,4

$$\bar{d} = \frac{13+7+9+5+3+2+1+0+6+1+4+3+2+6+12+4}{16} = \frac{78}{16}$$

$$= 4.88$$

d_i	$d_i - \bar{d}$	$(d_i - \bar{d})^2$	d_i	$d_i - \bar{d}$	$(d_i - \bar{d})^2$

13 7 9 5 3 2 1 0	8.12 2.12 4.12 0.12 -1.88 -2.88 -3.88 -4.88	65.93 4.494 16.92 1.44 3.534 8.294 15.05 23.81	6 1 4 3 2 6 12 4	1.12 -3.88 -0.88 -1.88 -2.88 1.12 7.12 -0.88	1.264 15.05 0.774 3.534 8.294 1.264 50.69 0.774
			78		221.166

$$s = \frac{221.166}{15} = 14.74$$

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{4.88}{\frac{14.74}{\sqrt{15}}} = 1.28$$

Conclusion: Since t calculated value 1.28 is less than t tabulated value 2.6 with 15 d.f and 1% level of significance we accept our null hypothesis H_0: hence the prescribed program is not effective .

Example 8: The daily wages in rupees of skilled workers in two cities are as follows.

City	size of sample	S.D of wages in the sample
City A	16	25
City B	13	32

Test at 5% level the equality of variances of the wage distribution in the two cities.

Solution: Set up the null hypothesis $H_0 : s_1^2 = s_2^2$

Alternative hypothesis $H_1^1 : s_1^2 \neq s_2^2$

Level of significance $\alpha = 0.05$

Test statistic to test the null hypothesis H0 is $F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$

Where $s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}{n_1 - 1}, s_2^2 = \frac{\sum_{j=1}^{n_2}(y_j - \bar{y})^2}{n_2 - 1}$

S.D of 1st sample is $\sigma_1 = 25$

$$\sigma_1^2 = (25)^2 = 625 = \sum_i^{n_1} \frac{(x_i - \bar{x})}{n_1} = \frac{s^2(n_1 - 1)}{n_1}$$

$$s_1^2 = \frac{\sigma_1^2 . n_1}{n_1 - 1} = \frac{16 \times (25)^2}{15} = 666.67$$

$$s_2^2 = \frac{\sigma_2^2 .. n_2}{n_2 - 1} = \frac{13 \times (32)^2}{12} = 1109.33$$

$$\therefore F = \frac{s_2^2}{s_1^2} \sim F(n_2 - 1, n_1 - 1)(\because taking greater value \text{ int } henumerator)$$

$$F = \frac{1109.33}{666.67} = 1.66$$

Conclusion: Since $F_{calculated}$ value 1.66 is less than $F_{tabulated}$ value $F_{0.05}(12,15)=2.48$ at 5% level of significance we accept our H_0:. Hence there is no significant differences between the variances.

Example 9: Five measurements of the output of two units have given the following results (in kilograms of material per one hour of operation) Assuming that both samples 5% level of significance if two population s have the same variance.

Unit A	14.1	10.1	14.7	13.7	14.0
Unit B	14.0	14.5	13.7	12.7	14.1

Solution let X_i-unit A

Y_i-unit B

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	y_j	$(y_j - \bar{y})$	$(y_j - \bar{y})^2$
14.1	0.8	0.64	14.0	0.2	0.04
10.1	-3.2	10.24	14.5	0.7	0.49
14.7	1.4	1.96	13.7	-0.1	0.01
13.7	0.4	0.16	12.7	-1.1	1.21
14.0	0.2	0.49	14.1	0.3	0.09
66.6		13.49	69.0		1.84

$$\bar{x} = \frac{66.6}{5} = 13.3, \bar{y} = \frac{69}{5} = 13.9$$

Null hypothesis $H_0 : s_1^2 = s_2^2$

Alternative hypothesis $H_1 : S_1^2 \neq s_2^2$

Level of significance $\alpha = 0.05$

Test statistic $F = \frac{s_1^2}{s_2^2} \grave{} F(n_1 - 1, n_2 - 1)$

$n_1 = 5, n_2 = 5.$

$$s_1^2 = \frac{\sum (x_i - \bar{x})}{n_1 - 1} = \frac{13.49}{4} = 3.37$$

$$s_2^2 = \frac{\sum (y_i - \bar{y})}{n_2 - 1} = \frac{1.84}{4} = 0.46$$

$$\therefore F = \frac{3.37}{0.46} = 7.3.$$

Conclusion: Since F calculated value 7.3 is greater than F- tabulated value F(4,4)=6.39 at 5% .level of significance we reject our null hypothesis H_0. Hence there is significant difference between the variances.

Example 10: it is desired to determine whether there is less variability in the silver plating done by company then in that in that done by company2.if independent random samples of size 12 of the two companies work yield s1=0.035 mil and s2=0.062 mil test the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ at the 0.05 level of significance.

Solution: set up the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$

Alternative hypothesis $H_1 : \sigma_1^2 < \sigma_2^2$

Level of significance $\alpha = 0.05$

Based on H1 the list is left tail test and at 5% level of significance with (11,11) d.f we have the F- table value 2.82.

Given $s_1^2 = (0.035)^2, s_2^2 = (0.062)^2$

Then the test statistic to test the null hypothesis is

$$F = \frac{s_2^2}{s_1^2} \sim F(n_2 - 1, n_1 - 1) d.f$$

$$\therefore F = \frac{(0.062)^2}{(0.035)^2} = 3.14$$

Conclusion: Since F- calculated value 3.14 is greater than F-tabulated value for (11,11) d.f at 5% .L.O.S is 2.82. Hence we reject our null hypothesis. Therefore the data support the conversion that the plating done by company 1 is less variable than that done by company2

Example 11: The following table shows the lives in hours of four batches of electric lamps

Batches

1	1600	1610	1650	1680	1700	1720
	1800					

2	1580	1640	1640	1700	1750	
3	1460 1740	1550 1820	1600	1620	1640	1660
4	1510	1520	1530	1570	1600	1680

Perform an analysis of variance of these data and show that a significance test does not reject their homogeneity.

Solution: We here set the null hypothesis

H_0: various batches are homogeneous.

Shifting the origin to 1640 and then dividing by 10 the above data reduces to

Batches		T_i	$\sum_i x^2{}_{ij}$
1	-4 -3 1 4 6 8 16 –	28	398
2	-6 0 0 6 11 - - -	11	193
3	-18 -9 -4 -2 0 2 10 18	-3	853
4	-13 -12 -11 -7 -4 4 - -	-43	515
	Total	-7	1,959

H0: the four batches of electric lamps are homogenous i.e., $\mu_1 = \mu_2 = \mu_3 = \mu_4$

R.S.S= $\sum_i \sum_j x^2{}_{ij} = 1959 and G = \sum_i \sum_j x_{ij} = -7, N = 26$

Correlation factor (C.F) = $\frac{G^2}{N} = \frac{(-7)^2}{26} = \frac{49}{26} = 1.8846$

TotalS.S=R.S.S – C.F

= 1959-1.8846=1957.1154

Between batches sum of squares $= \sum_{i=1}^{n} \left(\frac{T_i}{n_i} \right) - C.F$

$= \left[\frac{(28)^2}{7} + \frac{(11)^2}{5} + \frac{(-3)^2}{8} + \frac{(-43)^2}{6} \right] - C.F$

$= 445.4917 - 1.8846 = 443.6071$

With in (batches) sum of squares = total sum of squares – Between .S.S

=1957.1154-443.6071=1513.5083

ANOVA TABLE:

Source of variation	d.f	Sum of squares (S.S)	Mean sum of squares (M.S.S)	Variance F ratio

Between Batches Within Batches Error	3 22	443.6071 1573.5083	$\frac{443.6071}{3}=147.869$ $\frac{1513.5083}{22}=68.7958$	$F=\frac{147.869}{68.7958}=2.1493$
Total	25	1957.1154		

F- Table values $F_{0.05}$ for 3 and 22 d.f =3.05

$F_{0.01}$ for 3 and 22d.f =4.82

Conclusion: Since F-calculated value 2.1493 is < than F-tabulated values 3.05 (F0.05 for 3 and 22 d.f), 4.82 (F0.05 for 3 and 22 d.f) at both levels of significance we accept our null hypothesis . Hence we may conclude that the four batches to be homogeneous.

Example 12: Consider the results given in the following table for an experiment involving six treatments in four randomized blocks the treatments are indicated by numbers within brackets.

Yield for a randomized block experiment treatment and yield

Blocks 1	3	2	4	5	6	
1	24.7	27.7	20.6	16.2	16.2	24.9
	3	2	1	4	6	5
2	22.7	28.8	27.3	15.0	22.5	17.0
	6	4	1	3	2	5
3	26.3	19.6	38.5	36.8	39.5	15.4
	5	2	1	4	3	6
4	17.7	31.0	28.5	14.1	34.9	22.6

Test whether the treatments differ significantly. Also determine the critical difference between the means of any two treatments, and (ii) obtain the efficiency of this design relative to its layout as C.R.D

Solution: set up the null hypothesis H_t:$T_1=T_2=\ldots=T_6$

i.e the treatments as well as H_b: $6_1=6_2=6_3=6_4$

blocks are homogeneous . for finding the various sum of squares we rearrange the above table as follows.

blocks	Treatments	Block totals Bj	B_j
1 2 3 4	24.7 20.6 27.7 16.2 16.2 24.9 27.3 28.8 22.9 15.0 17.0 22.5 38.5 39.5 36.8 19.6 15.4 26.3 28.5 31.0 34.9 14.1 17.7 22.6	130.0 133.3 176.1 148.8	16,900 17768.89 31,011.21 22,141.41
Treatment	119.0 119.9 122.1 64.9 66.3	388.5=G	

totals T_i			
T_i^2	14161 14376.01 14908.41 4212.01 4395.69 9273.69		
Average	29.75 30.0 30.5 16.2 16.6 24.1		

Correction factor (C.F) = $\frac{G^2}{N} = \frac{346332.25}{24} = 14{,}430.51$

Raw sum of squares (Raw.S.S) = $\sum\sum y_{ij}^2 = 15{,}780.76$

Total sum of squares (T.S.S) = R.S.S-C.F=15,780.76-14430.51

=1,350.25

$$\frac{1}{6}\sum_{j=1}^{4} B_j^2 - C.F$$

Sum of squares due to blocks (S.S.B)= $= \frac{87899.63}{6} - 14{,}430.51$

$$= 219.43$$

Sum of squares due to treatments (S.S.T) = $\frac{1}{4}\sum_{i=1}^{6} T_i^2 - C.F$

$= \frac{61{,}326.81}{4} - 14{,}430.51 = 901.19$

Error sum of squares (E.S.S) = T.S.S – S.S.T – S.S.B

=1,350.25-901.19,-219.43=229.63

ANOVA TABLE

Source of variation	d.f	S.S	M.S.S	Variance (F) ration
Treatments	5	901.19	180.24	$F_t = \frac{180.24}{15.31} = 11.8$
Block	3	219.43	73.14	$F_t = \frac{73.14}{15.31} = 4.7$
Error	15	229.63	15.31	
Total	23	1,350.25		

F- table values $F_{0.05}(3.15) = 5.42, F_{0.05}(5.15) = 4.5$

Conclusion: Since F-calculated value due to treatment is 11.8 greater than the F- table value 5.42 at 5% level of significance we reject our H0 and we conclude that the treatment effects are not alike on the other hand f calculated value due to blocks is 4.7 is greater than f- table value 4.5 at 5% level of significance we reject our h0. We conclude that the blocks are not homogeneous.

(iii) since H_t: $T_1=T_2=\ldots.T_6$ is rejected , we are interested to find out which treatment means differ significantly c.o. for any two treatment means

= $t_{0.05}$ for error d.f X $\sqrt{\frac{2.s^2_E}{r}}$

Where 'r' is the number of times a treatment is replicated

$$= 2.131 \times \sqrt{\frac{2 \times 15.31}{4}} = 2.131 \times 2.8$$

$$= 5.97(approx)$$

By comparing the difference between the mean yields for difference treatments with the article difference, we find that the treatments 3,2 and 1 are alike in giving significantly high yields while treatments 4 and 5 are alike in significantly low yields by using the formula the efficiency of the above R.B.D relative to its layout as C.R.D is given by

$$E = \frac{r(t-1)SE^2 + (r+1)S^2_B}{(rt-1)S_e^2}$$

Where R is the number of replicates blocks and t is the number or treatments substituting the values, we get

$$E = \frac{4 \times 5 \times 15.31 + 3 \times 73.14}{23 \times 15.31} = 1.49$$

Hence gain in effecting is 49%

3.4 Exercise:

1. Studying the flow of traffic at two busy intersection between 4 pm and 6pm. It was found that on 40 week days there were on the average 247.3 cars approaching the first intersection from the south which made left turns , while on 30 week days there were on the average 254.1 cars approaching the section intersection from the south which made left turns. The corresponding sample standard deviations $\sigma_1 = 15.2, and, \sigma_2 = 18.7.$ test the null hypothesis $\mu_1 - \mu_2 = 0$ against the null hypothesis $\mu_1 \neq \mu_2$, at the level of significance =0.01.

2. Samples of students were drawn from two universities and from their weights in keg and standard deviations are calculated. Make a large sample test to test the significance of the difference between the means.

	Mean	S.D	Size of sample
University A	55	10	400
University B	57	15	100

3. A random sample of 400 men from one stage gives the mean pay of Rs. 200 per day with a S.D as 10 Rs. Another random sample or 400 men has a mean pay of rs.190 per day with as.D of Rs 9. Construct 99% .confidence interval for $\mu_1 - \mu_2$

Answer (8.26, 11.74)

4. Measurements of the fat content of two terms of ice-cream Brand A and Brand B yielded the following.

Brand A	13.5	14.0	13.6	12.9	13.0

Brand B	12.9	13.0	12.4	13.5	12.7

Test the null hypothesis $\mu_1 = \mu_2$ where μ_1, and, μ_2 are the respectively true average fat contents of the two kinds of icecream. Against the alternative hypothesis $\mu_1 = \mu_2$ at 5% level of significance

Ans : t=1.85,H0 is accepted)

5. The average losses of workers, before and after a certain program are given use 0.05 level of significance to test whether the program is effective (pained sample t – test). 40 and 35, 70 and 65, 45& 42, 120 & 116,35& 33, 55&50, 77&73.

Ans t=7.95,H0 is rejected.)

6. find the standard error of difference between the means and also find the confidence interval for difference of the means at 0.05 level for the following data.

	Size	mean	Standard deviation
Sample 1	9	69	4
Sample11	10	68	5

Ans: (S.E=2.7, C.I=(-4.7,6.7))

7. The daily wages in rupees of skilled workers in two cities are as follows :

city	Size of sample	S.D of wages in the sample
City A	16	25
City B	13	32

Test at 5% level of significance the equality of variances of the wage distribution in the two cities.

Ans : F=1.66, H0: accepted)

8. Below are given the gain in weights (inlbs) of pegs fed in two diets A and B.

Diet A: 25,32,30,34,24,14,32,24,30,31,35,25

Diet B: 44,34,22,10,47,31,40,30,32,35,18,21,35,29,22

Test, if the two diets differ significantly s regards their effect on increase in weight

Ans 1H=0.609, H_0 is accepted)

9. In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84.4 and in the other sample of 10 observations it was 102.6. Test whether this difference is significant at 5% level, given that the 5% point of F for n1=7 and n2=9 d.f is 3.29

Ans F= 1.057, H0 is accepted.

10. Two random samples gave the following results

Sample size sample mean sum of squares of deviations

			from the mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population at 5% level of significances

Given $F_{0.05}(9,11) = 2.90, F_{0.05}(11,9) = 3.10$
$\& \, t_{0.05}(20) = 2.086, t_{0.05}(22) = 2.07$

Ans : F-test : F=1.018 H0 accepted

T-test : t =0.742 H0 accepted.

11. Three processes A, B and C are tested to see whether their outputs are equivalent the following observations of output are made. Carryout the analysis of variances and state your conclusions.

A:	10	12	13	11	10	14	15	13
B:	9	11	10	12	13			
C:	11	10	15	14	12	13		

12. A test was given to five students taken at random from the fifth class of three schools of a town the individual scores are

School I:	9	7	6	5	8
School II:	7	4	5	4	5
School III:	6	5	6	7	6

Carryout the analysis of various and state your conclusions.

13. The following table gives the gains in weights of four different types of pigs fed on three different rations.

Test to see whether the rations or the pig types differ in their effect on mean weights.

Types of pigs

Types of rations

	I	II	III	IV
I	7	16	10	11
II	14	15	15	14
III	8	16	7	11

14. The following randomized block design with five treatments: O,A,B,C and D and 4 blocks was used the plan and yields in lbs. per plot were as follows

Block I	D67	B69	A70	C64	O65	
Block II		B71	C69	A73	O69	D68

Block III	O71	D70	C69	A75	B71
Block IV	C67	A70	O63	B69	D71

Prepare the analysis of variance table and test the homogeneity of means between treatments and blocks.

Also give the standard error between any two treatments means and between any two block means.

3.5 Summary:

In this unit an attempt is made to explain the tests of significance between two means for both large samples and small samples, for inferences concerning two variances, ANOVA one way classification and randomized complete block design. A number of examples are worked out and a good number of exercises are also given.

3.6 Technical Terms

Degrees of freedom

Critical difference

Analysis of variance

Randomized complete block design

Replication.

Unit-IV

Simple-Multiple LinearRegressionModels and Correlation

Syllabus: Simple linear regression, parameter destination inferences about stops intercepts, correlation, coefficient of delimitation, Multiple linear regression, least square procedures, A matrix approach, internal estimation.

Objective: This UNIT is prepared in such a way that after studying the student is expected to have a thorough compression of the above concepts like simple, multiple linear regression and correlation. Which are the important areas of investigationand statistical data analysis? The student will be having and well equipped with both theoretical as well as practical aspects of simple , multiple linear regression and correlation.

Structure of the UNITE;

4.1 Introduction:

Very often the Interest lies establishing the actual relationship between two or more variables. This Problem is dealt with regression. On the other hand, we are often not interested to know the actual relationship but or only interested in knowing the degree of relationship between two or more variables. This problem is dealt with correlation analysis. For both the studies, the number of variable may be two or more.

The concept of regression was first given by Sir FrancisGalton (1822-1911) in a study of inheritance of stature in the human being. To prove this biometrical fact, Real Pearson found the regression of son's height on father's height. But soon the use of regression technique became too common for a variety of problems. The relationship between variables, if it exists, may be linear or curvilinear.

But Scientific, social and economic Phenomena do not confine to two variables only. A large number of studies involve only more than two variables. In these studies we often need to give actual relationshipbetween three or more variables and /or to explain the strength of association between them. For instance, the cost of production of a manufactured product mainly depends on the cost of raw materials, the Labour charges and the cost of energy. The cost of a Crop mainly depends upon the cost of seeds, fertilizer,

irrigation, Pesticides and many farm operations In the both the examples, the cost of the Produced Productis a dependent factory, white others are independent factors.

We want to establish the relationship between the dependent variables; a mathematical equation can be given to do this. This type of mathematical equation is known as a mathematical model. The equation pertaining to such a relationship may be of any type. But we deal with a linear relationship which represents a plane according to the number of variables involved.

4.2. Simple Linear Regression &Correlation:

The term “regression” literally means “Stepping back towards the average. If there exists some relation between two variables, their scatter diagram shall be having points clustering near about some curve. If this curve. If this curve is a straight line, it suggest some linear relationship between the variables and this straight line is known as the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of “best fit” and is obtained by the principles of least squares.

LINES ofRegressionY on X and X and Y:

Let us Suppose that in the vicariate distribution (xi,yi): i =1,2,----n: y is dependent variable and X is independent variable. Let the line of regression of Y on X be Y= a+bx According to the principle of least squares the normal equations for estimating a and b are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \text{ ----------(1)}$$

And

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \text{ --------(2)}$$

From (1) on dividing by n, we get

$$\bar{y} = a + b\bar{x} \text{ ----------(3)}$$

Thus the line of regression of y on x passes through the point $(\bar{x}, \bar{y})$Now

$$\mu_{11} = \text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y} \text{ -------(4)}$$

Also

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_x^2 + \bar{x}^2 \text{ -----(5)}$$

Dividing (2) by n and using (4) and (5), we get

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_x^2 + \bar{x}^2) \text{ ----------(6)}$$

Multiplying (3) by $\bar{x}$ and then subtracting from (6), we get

$$\mu_{11} = b\sigma_x^2 \Rightarrow b = \frac{\mu_{11}}{\sigma_x^2} --------(7)$$

Since b is the slope of the line of regression of y on x and since the line of regression passes through the point ($\bar{x}, \bar{y}$), its equation is

$$y - \bar{y} = t(x - \bar{x}) = \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x}) ------------(8)$$

$$y - \bar{y} = r = \frac{\sigma_y}{\sigma_x}(x - \bar{x}) ------------(9)$$

$$\because r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x . \sigma_y}$$

By interchanging the variables

X and Y in (8) and (9) the equation

Of the line of regression of X on Y becomes

$$X - \bar{X} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y}) ----------(10)$$

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{y}) ----------(11)$$

Note: we always have two lines of regression except in particular case of perfect correlation when both lines coincide, we get only one line. That is, in case of perfect correlation, ($r = \pm 1$), both the lines of regression coincide.

Regression coefficients:

'b', the slope of the line of regression of Y and X is also called the coefficient of regression of Y on X. it represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. then we write.

$$b_{yx} = \text{Re}\,gression \text{ coefficient of yon x} = \frac{\mu_{11}}{\sigma_x^2} r\frac{\sigma_4}{\sigma_x} ----(1)$$

Similarly, the coefficient of regression of X and Y indicates the change in the value of variable x Corresponding to a Unit change in the value of variable y and is given by b x y =

Regression coefficient of X on Y = $\frac{\mu_{11}}{\sigma_y^2} = r\frac{\sigma_x}{\sigma_y}$ ---------(2)

Properties :

1. Correlation Coefficient is the geometric mean between the regression coefficients. i.e.,
   $r = \pm\sqrt{b_{xy}} \times b_{yx}$
2. If one of the regression coefficients is greater than unity, the other must be less than unity.
   $b_{xy} \le \frac{1}{b_{yx}} < 1$
3. Arithmetic mean of the egression coefficients is greater than the correlation coefficient , protected
   $\frac{1}{n}(b_{yx} + b_{xy}) \ge r$
4. Regression coefficients are independent of the change of origin but not of scale.

$b_{yx} = \frac{k}{h} b_{vu}$

Note: Angle between two lines of Regression is

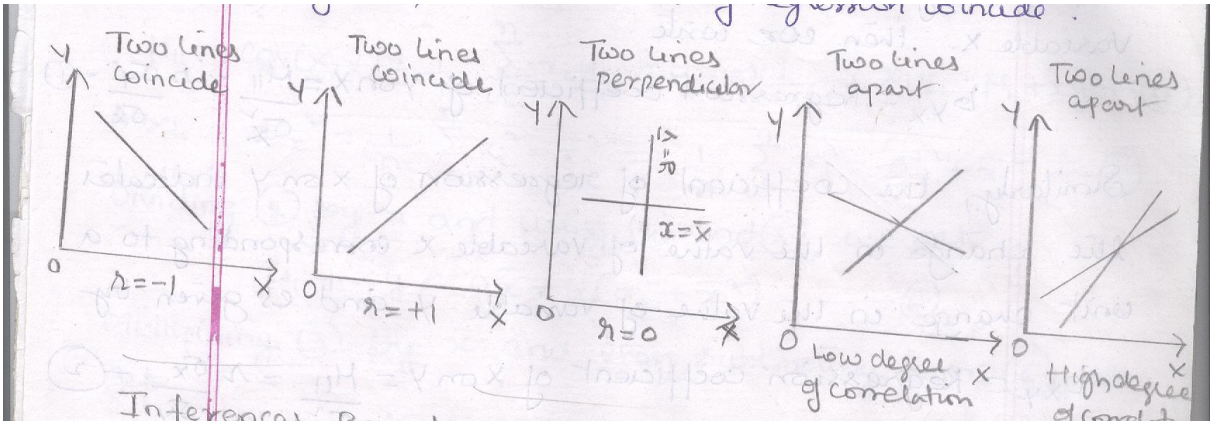
$$\theta = \tan^{-1}\left\{\frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\}$$

Case: $r = 0, \tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Thus if the two variables re uncorrelated, the lines of regression become perpendicular to each other.

If $r = \pm 1, \tan\theta = 0 \Rightarrow 0 or \pi$

In this case the two lines of regression either coincide or they are parallel to each other. But since both the lines of regression pass through the point ($\bar{x}$, $\bar{\bar{y}}$), they cannot be parallel. Hence in the case of perfect correlation, positive or negative, the two lines of regression coincide.



The method of the proceeding section is used when the relationship between x and the mean of y is linear or close enough to a straight line so that the least- squares line yields reasonably good predications. We shall assume that the regression is linear and further more that the n random variables Yi (i= 1,2,---n) are independently normally distributed with the means and the common variance . Equivalently, we write the model as

$$y_i = \alpha + \beta x_i + E_i, i = 1,2.....n(1)$$

Where it is assumed that the areindependently normally distributed random variables having zero means and the common variance σ^2.

We know that the normal equations to fit a straight line y= a+ bx by least squares method are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \text{ -------(2)}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \text{ ------(3)}$$

The following expressions pertaining to the sample values (xi, yi) occur so often that can written as

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n \text{ -------(4)}$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n \text{ -------(5)}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \Big/ n \text{ -------(6)}$$

Dividing the equation (2) by ‘n’ we get

$$\sum_{i=1}^{n}\frac{y_i}{n}=a+b\sum_{i=1}^{n}\frac{x_i}{n}$$

$$\bar{y}=a+b\bar{x}$$

$$a=\bar{y}-b\bar{x}-------(7)$$

Substituting the value of ain (3) we get

$$\sum_{i=1}^{n}x_i y_i=(\bar{y}-b\bar{x})\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2$$

$$\sum_{i=1}^{n}x_i y_i-\bar{y}\sum_{i=1}^{n}x_i=b\left(\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i\right)$$

$$\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}\frac{\bar{y}_i}{n}\sum_{i=1}^{n}x_i=b\left(\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2\right)$$

$$\frac{\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}\frac{\bar{y}_i}{n}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}$$

$$\frac{\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}\frac{\bar{y}_i}{n}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}$$

$$=\frac{S_{xy}}{S_{xx}}$$

From equation 4 & 6

$$b=\frac{S_{xy}}{S_{xx}},\&a=\bar{y}-b.\bar{x}$$

There is a relation between Sxx and Syy and the respective sample variances of the x’s and y’s

$$S_X^2 = \frac{S_{xx}}{n-1}, S_y^2 = \frac{S_{yy}}{n-1}$$

The estimate of σ^2 is

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (a+bx_i)\right)^2$$

Where is the standard error of estimate. Also, the sum of squares given by (n-2) refers as the residual sum of squares or the error sum of squares. An equivalent formula for this estimate of is

$$= \frac{S_{YY} - \frac{(S_{xy})^2}{S_{XX}}}{n-2}$$

In the above formula (n-2) is used to make the estimator for unbiased. The two regression coefficients α and β had to be replaced by their least square estimates. Therefore the degrees of freedom is (n-2) in the formula from equation (1) and under these assumptions

Is a value of a random variable having the X^2 = distribution with (n-2) degrees of freedom.

Hence under the assumptions the statistics for inference about α and β is

$$t = \frac{(a-\alpha)}{S_e}\sqrt{\frac{nS_{XX}}{S_{XX} + n(\bar{x})^2}} \text{ -----(8)}$$

$$t = \frac{(b-\beta)}{S_e}\sqrt{S_{XX}} \text{ -----(9)}$$

Are the values of random variables having the t- distribution with (n-2) degrees of freedom

ConfidenceLimitsforRegression coefficients:

To construct confidence intervals for the regression coefficients α and β, we substitute for the middle term of $-t_{\alpha/2} < t < t_{\alpha/2}$ are the appropriate t statistic of equation (8) and (9) we get confidence limits for regression coefficients

For $\alpha = a \pm t_{\alpha/2}.S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

And for $\beta : b \pm t_{\alpha/2}.S_e\frac{1}{\sqrt{S_{xx}}}$

Estimating α +βx :

If x is fixed say x_0, the quantity we want to estimate is $\alpha + \beta x_0$, where a and b are obtained for a straight line y=a+ bx by the method of least squares. This estimator is unbiased and the variance is

$$\sigma^2 \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

And $(1-\alpha)100\%$ confidence interval for $\alpha + \beta x_0$ is

$$(a + bx_0) \pm t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Where the number of degrees of freedom for is n-2

Prediction of Future value of y:

we can find the confidence interval for the prediction of a future value of y whenx= x_0, where x_0 is within the range of experimentation.

It y is a random variable having a normal distribution with the mean $\alpha + \beta x_0$ and the variances σ^2 is a random variable having a normal distribution with zero mean and the variance σ^2)

If α and β are not known, we can consider y-a-b x_0 then confidence interval for the prediction of a future value of y when x= x_0

$$(a + bx_0) \pm t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Correlation:

Regression technique provides the actual relationship between two or more variables. But the Scientists are not always interested in this linear or curvilinear relationship. Often the interest lies only in knowing the extent of inter dependence between two or more variables. In this situation, correlation methods serve our purpose. If the two variables say x and y are linearly related they are said to be correlated. The correlation between two variables is also known as simple correlation. The measure of correlation was given by Prof. Karl Pearson in 1896 in the form of correlation coefficient.

Correlation Definition:

The relationship between the two variables such that a change in the one variable results in a positive or negative change in the other and also greater change in one variable results in a corresponding greater change in the other is called a correlation.

For a change in one variable, there is a corresponding change in the other variable. The variables are said to be correlated.

i. Of the two variables deviate in the same direction, the correlation is said to be direct or positive.

ii. If the variables deviate in the opposite direction the correlation is said to be inverse or negative.

iii. If the change in one variable corresponds to a proportional change in the other variable thenthe correlation is said to be perfect.

Karl Pearson coefficient of correlation:

Measure of intensity or degree of linear relationship between two variables is called correlation coefficient and it is devoted by

Coefficient of correlation $r = \frac{S_{XX}}{\sqrt{S_{XX} S_{YY}}}$

Where $S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$

$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

$\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sigma_x \sqrt{n}$

$\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \sigma_y . \sqrt{n}$

There fore r can be written as $\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n..\sigma_x.\sigma_y}$

$$\frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Note:

i) The coefficient of correlation lies between -1 and 1.

ii) If =0, the variables are not correlated or there is no correlation between the variables x and y.

iii) If =1, we say that the variables are positively, perfectly correlated.

iv) If =-1, we say that the variables are negatively, perfectly correlated.

Fisher z- transformation:

If the value of is based on a random sample from a bivariate normal population, we can perform a test of significance for $H_{O:}$ $P=P_O$, or construct a confidence interval for Point the basis of the transformation

$$z = \frac{1}{2}\log\frac{1+r}{1-r}$$

Where p=$1-\frac{\sigma^2}{\sigma_2^2}$

If P =0, the two random variables are independent

σ^2 = measure of the variation of y

σ_2^2 = measure of variation of y when x is unknown.

$\sigma_2^2 - \sigma^2$ = measure of variation of y that is accounted by the linear relationship with x.

P^2= proposition of the variation of y that can we attributed to the linear relationship with x.

P= population correlation coefficient statistic for inference about P is

$$Z = \frac{Z-\mu_z}{\frac{1}{\sqrt{n-3}}} = \frac{\sqrt{n-3}}{2}.l_n\frac{(l+r)(1-p)}{(l+r)(1+p)}$$

Statistic for test of

$$Z = \sqrt{Z-3Z} - \frac{\sqrt{n-3}}{2}l_n\frac{(1+r)}{(1-r)}$$

Confidence interval for μ_z

$$= Z - \frac{Z_{\alpha/2}}{\sqrt{n-3}} < \mu_2 < Z + \frac{Z_{\alpha/2}}{\sqrt{n-3}}$$

Coefficient of determination:

In regression analysis we estimate the value of y for a given value of x. the estimate is good if the mean sum of squares due to error is minimum. We know the error mean sum of squares is

$$\sigma_e^2 = \sigma_y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} = \sigma_Y^2\left(1 - \frac{\sigma_{XY}^2}{\sigma_X^2\sigma_Y^2}\right) = \sigma_Y^2\left(1 - p^2\right)$$

If p=0, there is no purpose of regression of y on x as $\sigma_e^2 = \sigma_y^2$. if P is high, will be small and y can be determined accurately through x. the quantity P^2 is called the coefficient of determination and $(1-P^2)$, the coefficient of non determination. Also the quantity $\sqrt{1-p^2}$ is called the coefficient of alienation.

4.3. Multiple Linear Regression Models:

Least Square Procedures for Model Fitting**:**

Having collected some data, it is desirable to find out the form of universe of which the observed values are regarded as a sample. In other words, we try to find a functional relationship between the observed values so as to have a clear picture of the universe of which our observations are a past. If is neither necessary nor possible that all the observed values should strictly satisfy this relationship, but the curve, representing this relationship, should as far as possible pass closely to all the points. The difference between the observed values and expected values is known as residual and the task is to minimize these residuals. Since these differences may be positive in some cases and negative in others; it is more convenient to make the sum of squares of these residuals a minimum. This is known as the method of least squares.

Fitting of a Straight line:

Let us consider the fitting of a straight line y=a+ b x _____ (1)

To a set of n points (xi, yi);i=1, 2, ------n. equation represents a family of straight lines for different value of the arbitrary constants ‘a’ and ‘b’. The problem is to determine ‘a’ and ‘b’ so that the line (1) is the line of best fit. According to the principle of least squares we have to determine a and b so that

$$E = \sum_{i=1}^{n}\left(y_i - a - bx_i\right)^2$$

Is minimum from the principle of maxima and minima the partial derivatives of E,(w.r.t.) a and b should vanish separately. i.e .,

$$\frac{\partial E}{\partial a} = 0 \Rightarrow -2\sum_{i=1}^{n}\left(y_i - a - bx_i\right) and\ \frac{\partial E}{\partial f} = 0 = -2\sum_{i=1}^{n} x_i\left(y_i - a - bx_i\right) \text{------(2)}$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i and \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \text{-----(3)}$$

Equations (2) and (3) are known as the normal equations for estimating a and b.

All the quantities $\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i and \sum_{i=1}^{n} x_i y_i$

Can be obtained from the equations (3) can be solved for a and b. with the values of a and b so obtained equation (1) is the line best fit to the given set of points (xi, yi); i= 1, 2, ----- n.

Note: the equation of the line of best fit of y and x is obtained on eliminating a and b in (1) and (3) and can be expressed in the determinant form as follows.

Fitting of second degree parabola

Let $y= a+b+cx^2$ _____ (1)

Be the second degree parabola of best fit to set of n points (xi, yi); i= 1, 2, ----- n.the principles of least squares, we have to determine a, b and c. so that

$$E = \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right)^2$$

Is minimum. Equating to zero the partial derivatives of E with respect a,b and c separately, we get the normal equations for estimating a,b and c as

$$\frac{\partial E}{\partial a} = 0 = -2\sum \left(y_i - a - bx_i - cx_i^2\right)$$

$$\frac{\partial E}{\partial a} = 0 = -2\sum x_i \left(y_i - a - bx_i - cx_i^2\right) \text{----------(2)}$$

$$\frac{\partial E}{\partial a} = 0 = -2\sum x_i^2 \left(y_i - a - bx_i - cx_i^2\right)$$

$$\sum y_i = na + b\sum x_i + c\sum x_i^2$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \text{ ----------(3)}$$

$$\sum x_i^2 y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4$$

Summation taken over I from 1 to n.

For given set of points (xi, yi); i= 1, 2, ----- n. equations 93) can be solved for a, b and c and with these values of a, b and c equation (1) is the parabola of best fit.

Note: Eliminating a, b and c in (1) and (3) the parabola of best fit of y on x is given by

Fitting of a Power curve y=a x b

Fitting of a power curve y=a x b _____ (1)

To a set of n points. Taking logarithan on both sides

Weset log y = log a + b log x

$\Rightarrow U = A + bV$

$U = \log y, A = \log a, and V = \log X$

This is a linear equation in V and U

Normal equations for estimating A and B are

$\sum U = nA + b\sum V and \sum UV = A\sum V + b\sum V^2 \text{ – – – – –} (2)$

The equation (2) can be solved for A and b and consequently we set a= antilog A. with the values of a and b so obtained (1) is the curve of best fit to the set of n points.

Fitting of Exponential curves

To a set of n points.

i) $y = ab^X \text{ – – – –} (1)$

Taking log on both sides, we get

$\log y = \log a + X \log b \Rightarrow U = A + BX$

$where, U = \log y, A = \log a, and, B = \log b.$

This is linear equation in X and U. the normal equations for estimating A and B, we finally get a= antilog A and b are

$\sum U = nA + B\sum X, and, \sum XU = A\sum X + B\sum X^2$

Solving these equations for A and B, we finally get

A=antilog and b= antilog.

With the values of a and b (1) is the curve of best fit to the given set of n points.

$y = ae^{bx} \text{ – – – – – –} (1)$

Taking log on both sides we get

Logy=loga+bXloge=loga+(bloge)X

U=A+BX

Where U=logy, A=loga and B=blogeX

This is linear equation in X and U

Thus the normal equations are

$$\sum U = nA + B\sum X, and, \sum XU = A\sum X + B\sum X^2$$

From these we final A and B and consequently

$$a = anti\log(A) and, b = \frac{B}{\log e}.$$

Fitting of Multiple Regression Model:

In multiple regressions, we deal with data consisting of n(r+1)touples $(x_{1i}, x_{2i}.....x_{ri}y_i)$, where the x's are again assumed to be known without error while the y' s are values of random variables. Data of this kind arise, for example, in studies designed to determine the effect of various climatic conditionson a metal's resistance to corrosion; the effect of kiln temperature, humidity and iron content on the strength of a ceramic coating; or the effect of factory production, consumption level and stocks in storage on the piece of a product.

As in the case of one independent variable, we shall first consider the problem where the regression equation is linear, namely, where for any given set of values $x_1, x_2, ...$ and x_r , for the (r) independent variables, the mean of the distribution of y is given by

$$a_0 + a_1x_1 + a_2x_2 + ...a_rx_r.....$$

For two independent variables, the problem of fitting a plane to a set of n points with coordinates (x_{i1}, x_{i2}, y_i). Applying the method of least squares estimates of the coefficients a_0, a_1, and, a_2, we minimize

$$y = a_0 + a_1x_1 + a_2x_2 + ...a_rx_r.....(1)$$

Let the residual be E_i

$$E_i = [y_i - a_0 + a_ix_{1i} + a_2x_{2i} + ... + a_rx_{ri}] - - - - - -(2)$$

$$\sup pose, S = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} [y_i - a_0 + a_ix_{1i} + a_2x_{2i} + ... + a_rx_{ri}] - - - - - -(3)$$

The first order particular deviators of S with respect to $a_0, a_1, and, a_2 .....a_r$ should vanish.

$$\therefore \frac{\partial s}{\partial a_0} = 0, \frac{\partial s}{\partial a_1} = 0, \frac{\partial s}{\partial a_2} = 0, \frac{\partial s}{\partial a_3} = 0, ... \frac{\partial s}{\partial a_r} = 0 - - - - - (4)$$

$$-2\sum_{i=1}^{n} [y_i - (a_0 + a_ix_{1i} + a_2x_{2i} + ... + a_rx_{ri})] = 0$$

$$-2\sum_{i=1}^{n} x_{1i}[y_i - (a_0 + a_ix_{1i} + a_2x_{2i} + ... + a_rx_{ri})] = 0$$

$$-2\sum_{i=1}^{n} x_{2i}\left[y_i-(a_0+a_i x_{1i}+a_2 x_{2i}+.....+a_r x_{ri})\right]=0 \text{ ----------(5)}$$

$$-2\sum_{i=1}^{n} x_{ri}\left[y_i-(a_0+a_i x_{1i}+a_2 x_{2i}+.....+a_r x_{ri})\right]=0$$

$$\sum_{i=1}^{n} y_i = na_0 + a_1\sum_{i=1}^{n} x_{1i} + a_2\sum_{i=1}^{n} x_{2i} + ... + a_r\sum_{i=1}^{n} x_{ri}$$

$$\sum_{i=1}^{n} x_i y_i = a_0\sum_{i=1}^{n} x_{1i} + a_1\sum_{i=1}^{n} x_{1i}^2 + a_2\sum_{i=1}^{n} x_{1i}x_{2i} + ... + a_r\sum_{i=1}^{n} x_{2i}x_{ri}$$

$$\sum_{i=1}^{n} x_i y_i = a_0\sum_{i=1}^{n} x_{2i} + a_1\sum_{i=1}^{n} x_{1i}x_{2i} + a_2\sum_{i=1}^{n} x_{2i}^2 + ... + a_r\sum_{i=1}^{n} x_{2i}x_{ri}$$

$$\sum_{i=1}^{n} x_i y_i = a_0\sum_{i=1}^{n} x_{2i}x_{ri} + a_1\sum_{i=1}^{n} x_{ri}x_{1i} + ... + a_r\sum_{i=1}^{n} x_{ri}^2 \text{ ----------(6)}$$

The set of equation (6) are normal equation to fit

$y = a_0 + a_1x_1 + a_2x_2 + ...a_r x_r$

Suppose $y = a_0 + a_1x_1 + a_2x_2 + ...a_r x_r$, $x_1, and x_2$ are two independent variables.

The normal equations are

$$\sum_{i=1}^{n} y_i = na_0 + a_1\sum_{i=1}^{n} x_{1i} + a_2\sum_{i=1}^{n} x_{2i}$$

$$\sum_{i=1}^{n} x_i y_i = a_0\sum_{i=1}^{n} x_{1i} + a_1\sum_{i=1}^{n} x_{1i}^2 + a_2\sum_{i=1}^{n} x_{1i}x_{2i}$$

$$\sum_{i=1}^{n} x_{2i} y_i = a_0\sum_{i=1}^{n} x_{2i} + a_1\sum_{i=1}^{n} x_{1i}x_{2i} + ... + a_r\sum_{i=1}^{n} x_{2i}x_{ri}$$

Multiple Linear Regression a Multiple approach to least square

The model that we are using in multiple linear regressionslends itself uniquely to a unified treatment in matrix notation. In order to express the normal equations in matrix notation, let us define the following three matrices.

$$X = \begin{bmatrix} \vdots & x_{11} & x_{12} \\ \vdots & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ \vdots & x_{n1} & x_{n2} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} and\ a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

The first one X, is an n x (2=1) matrix consisting essentially of the given values of x's, with column of 1's appended to accommodate the constant term. Y is an n x 1 matrix (orcolumn vector) consisting of observed values of the response variable. And b is the (2+1) x 1 matrix. (Or column vector) consisting of the least squares estimates of the regression coefficients. The least squares estimates of the multiple regression coefficients are given by

$$b = (X^1 X)^{-1} X^1 Y$$

Where X^1 is the transpose of x and $(X^1 X)^{-1}$ is the inverse of .to verify this relation, we first determine $X^1 X, X^1 \times b, and, X^1 Y$

$$X^1 X = \begin{bmatrix} n & \Sigma x_{1i} & \Sigma x_{2i} \\ \Sigma x_{1i} & \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{2i} & \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} = \begin{bmatrix} n & \Sigma x_{1i} & \Sigma x_{2i} \\ \Sigma x_{1i} & \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{2i} & \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}$$

$$X^1 * b = \begin{bmatrix} a_0 n + a_1 \Sigma x_{1i} + a_2 \Sigma x_{2i} \\ a_0 \Sigma x_{1i} + a_1 \Sigma x_{1i}^2 + a_2 \Sigma x_{1i} x_{2i} \\ a_0 \Sigma x_{2i} + a_1 \Sigma x_{2i} x_{1i} + a_2 \Sigma x_{2i}^2 \end{bmatrix}$$

$$X^1 y = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{1i} y_i \\ \Sigma x_{2i} y_i \end{bmatrix}$$

Indentifying the elements of $X^1 \times b$ as the expression on the right hand side of the normal equation and those of $X^1 y$ as the expression on the left hand side, we can write.

$$X^1 \times b = X^1 y$$

Multiplying on the lift by $X^1 X)^{-1}$, we get

$$(X^1 X)^{-1} X^1 \times b = (X^1 X)^{-1} X^1 y$$

And finally $b = (X^1 X)^{-1} X^1 y$ equals the (2+1) X (2+1) identity matrix I, and by definition if =b. we have assumed here that X^1X is non singular , so that its inverse exists.

Interval Estimation or Confidence Intervals in Multiple Regression:

Under the normality assumption,

$$\hat{\beta} \sim MN(\beta, \sigma^2 (X^1 X)^{-1})$$

$$\text{cov}(\beta) = \hat{\sigma}^2 (X^1 X)^{-1} = \sigma^2 C$$

$Var(\hat{\beta}_j) = \sigma^2 C_{jj}$

Standard error : $Se(\beta_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$

$\beta_j \in \hat{\beta}_j \pm t_{\alpha/2}, n-p \sqrt{\hat{\sigma}^2 C_{jj}}$

4.4. Worked out Examples:

Example 1:

The following data pertain to the marks in subjects A and B in a certain examination.

Mean marks in A = 39.5, Mean marks in B = 47.5 Standard deviation of marks in A = 10.8, Standard deviation of marks in B = 16.8, coefficient of correlation between marks in A and marks in B = 0.42, Find the two egression lines. Find the marks in B for candidates who secured 50 marks in A.

Solution:Given Mean of A = $\bar{x}$ = 39.5

Mean of B = y = 47.5

Standard deviation of marks in A = 10.8 = σ_x

Standard deviationof marks in B = 10.8 = σ_y

Coefficient of correlation between the marks in A and B = 0.42

The line of regression of y on x is

$$(y - \bar{y}) = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$$

i.e., $(y - 47.5) = 0.42\frac{16.8}{10.8}(x - 39.5)$

$y = 0.651 + 21.82$

The line of regression of x and y

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$(x-39.5)=0.45\frac{10.8}{16.8}(y-47.5)$$

i.e., $x=0.27y+26.67$

$when x=50$

$y=0.651\times 50+21.82=54.37$

Example 2:

Twenty five pairs of value of vitiates x and y led to the following results . a subsequent scrutiny showed that two pairs of values were copied down as

x	y
8	18
8	6

x	y
8	12
6	8

Find correct value of r and correct lines of regression solution: In correct $\sum x$=127

The values of incorrect values 8 and 8, sum=16 correct values 8 and 6, total =14

$\sum x$should be reduced by 2

Correct$\sum x$= 127-2=125

Incorrect y values 14 and 6 their sum = 20

Correct values of y 12 and 8 their sum =20

Hence there is no difference in $\sum y$

Correct$\sum y$ = 100

$$n=25, then, \bar{x}=\frac{\sum x_i}{n}=\frac{125}{25}=5$$

$$\bar{y}=\frac{\sum y_i}{n}=\frac{100}{25}=4$$

Incorrect $x_1^2+x_2^2=8^2+8^2=64+64=128$

Correct $x_1^2+x_2^2=8^2+6^2=64+36=100$

$\therefore \sum x^2$ should be reduced by 28.

Incorrect $y_1^2+y_2^2=196+36=232$

Correct $y_1^2+y_2^2=144+64=208$

Hence $\sum y^2$ should be reduced by 24.

$\therefore correct \sum y^2=449-24=425$

$$r=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}=\frac{\frac{1}{n}\sum xy-\bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2\frac{1}{n}\sum(y_i-\bar{y})^2}}$$

$$\sigma_x^2=\frac{\sum x_i^2}{n}-\bar{x}^2=\frac{732}{25}-25=29.28-25=4.28$$

$$\sigma_x=\sqrt{4.28}=2.07,\sigma_y^2=\frac{425}{25}-16=1$$

Then $\sigma_y=1$

Incorrect $\sum xy: x_1y_1+x_2y_2=8\times14+8\times6=112+48=160$

Correct $\sum xy=8\times12+6\times8=96+48=144$

$\therefore \sum xy$ should be reduced by 16

Hence correct $\sum xy$= 500-16=484

Correlation coefficient $r=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}=\frac{\frac{1}{n}\sum xy-\bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2\frac{1}{n}\sum(y_i-\bar{y})^2}}$

$$=\frac{\frac{1}{n}\sum xy-\bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum(x^2-\bar{x}^2)\frac{1}{n}\sum(y^2-\bar{y}^2)}}$$

$$=\frac{\frac{1}{25}\times485-4\times5}{\sqrt{\left(\frac{0.732}{25}-25\right)\left(\frac{425}{25}-25\right)}}=\frac{-0.64}{2.07}=-0.31$$

The line of regression of y on x

$$(y-\bar{y})=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$y-4=-0.31\frac{1}{2.07}(x-5)$$

$$y=-0.15x+4.75$$

The line of regression of x on y is

$$x-5=-0.31\left(\frac{2.07}{1}\right)(y-4)$$

$$x=0.6417y+7.57$$

Note: 1. Standard error of $r=\frac{1-r^2}{\sqrt{n}}$, where r is coefficient of correlation and n is the sample size.

2.Probable error of = 0.6745× S.E(r)

Example 3:

Find the standard error and probable error where and n given by 0.31, 25

Solution: given r=-0.31, n = 25.

Standard of $r=\frac{1-r^2}{\sqrt{n}}=\frac{1-0.961}{\sqrt{25}}=\frac{0.039}{5}=0.008$

Probable error = 0.6745×S.E.(r) = 0.6745×0.008=0.0054.

Example: if r=0.75, n=16, find the standard error and probable error.

Solution: Standard error of $r=\frac{1-r^2}{\sqrt{n}}=\frac{1-(0.75)^2}{\sqrt{16}}=\frac{1-0.5625}{4}$

$$=\frac{0.4375}{4}=0.1094\left[\because r=0.75, n=16\right]$$

Probable error = 0.6745×S.E (r) =0.6745× 0.1094=0.07378

Example 4:

The following results were obtained in the analysis of data on yield of dry bark in ounces Y and age in years X of 200 cinchona plants.

Average:	x	y
	9.2	16.5
Standard deviation:	2.1	4.2

Correlation coefficient r = 0.84

Find the two lines of regression and estimate the yield of dry bark of a plant of age 8 years.

Solution: given average of x i.e., $\bar{x}=9.2$

average of y i.e., $\bar{y}=16.5$

Standard deviation of x i.e., $\sigma_x=2.1$

Standard deviation of y i.e., $\sigma_y = 4.2$

Coefficient of correlation r = 0.84

The line of regression y and x is

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \Rightarrow (y - 16.5) = 0.84(\frac{4.2}{2.1})(x - 9.2)$$

$$y - 16.5 = 1.68x - 15.456 \Rightarrow y = 1.68x + 1.044$$

The line regression of x on y is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \Rightarrow x - 9.2 = 0.84\left(\frac{2.1}{4.2}\right)(4 - 16.5$$

$$ie. X = 0.42y - 6.930 + 9.2$$

$$x = 0.42y + 2.270$$

When x = 8 years

Y = 1.68×8+1.044=14.48

$\therefore$ The yield of dry bark is 14.48

Example 5:

if x = 2y+3 and y= kx+6 are the regression lines of x on y and y on x respectively. Then (i) show that $0 \le k \le 1/2$ (ii) if k=1/8, find r and $(\bar{x}, \bar{y})$.

Solution: given regression line x and y is x= 2y+3.

given regression line y and x is y= kx+6.

i) Then the regression coefficient of x on y is b_{xy}= 2
The regression coefficient of y on x is b_{yx} = k
$r^2 = b_{yx}.b_{xy} = 2k$

We know that $0 \le r^2 \le 1$

$$\therefore 0 \le 2k \le 1 (\because r^2 = 2k)$$

$$\Rightarrow 0 \le k \le \frac{1}{2}$$

ii) If k = 1/8. The regression line y on x
$\Rightarrow$y=1/8x+6

i.e., 8y= x+48 or 8y-x=48 _____ (1)
The regression line x on y

X = 2y +3 or x-2y=3 _____ (2)

Solution (1)&(2) we get -x+8y=48

x -2y = 3

$$6y = 51 \Rightarrow y = \frac{51}{6} = 8.5$$
$$x = 2(8.5) + 3 = 17 + 3 = 20$$
$$then, (\bar{x}, \bar{y}) = (20, 8.5)$$
$$\therefore r^2 = b_{xy}.b_{yx} = 2k = 2.\frac{1}{8} = \frac{1}{4}$$
$$\therefore r = \frac{1}{2} = 0.5$$

Hence the coefficient of correlation is r=0.5

Example 6:

For an army personnel of strength 25, the regression of weight of kidneys (y) on weight of heart 9x0 both measured in ounces is y- 0.399x=0.934 and the regression of weight of heart (X) on weight of kidney (y) is x-1.212y+2.461=0.
Find the coefficient of correlation between x and y and their mean values.

Solution: the regression coefficient y on x is

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.399$$

Since the regression line y on x is y-0.399x=6.934
The line of regression x on y is x=10212y- - 2.461

$\therefore$ the regression coefficient x on y is $b_{xy} = r\frac{\sigma_x}{\sigma_y} = 1.212$

$$r^2 = b_{yx}.b_{xy} = 0.399 \times 1.212 = 0.484$$
$$\therefore r = 0.7$$

Since the two lines of regression intersect at $(\bar{\chi}, \bar{y})$

$\therefore$ we get $(\bar{\chi}, \bar{y})$ by solving the two lines of regression

$$y - 0.399\times = 6.934$$
$$= \frac{-0.4844 + 0.399x = 0.982}{}$$

$$0.516y = 7.916$$

$$y = \frac{7.916}{0.516} = 15.34$$

$$x = 1.212(15.34) - 2.461 = 16.13$$

Example 7:

The equations of two regression lines obtained in a correlation analysis are 3x+12y=19, 3y+9x=46 find (i) coefficient of correlation. (ii) Mean values of x and y and (iii) the ratio of the coefficient of variability of x to that of y

Solution:

(i) given 3x+12y=19 line of regression y on x regression coefficient of y on x

Regression coefficient of y on x is $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

i.e., 12y=19-3x, $y = \frac{19}{12} - \frac{3}{12}x$

$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{-3}{12} = -0.25$

Line of regression x on y is 3y+9x=46

Then 9x=46-3y $\Rightarrow$ x=46/9 - 3/9y

i.e., the regression coefficient x on y is

$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{-3}{9}$

$\therefore r^2 = b_{xy}b_{yx} = \frac{-3}{12}.\frac{-3}{9} = 12$

$\therefore r = \frac{-1}{2\sqrt{3}}$

since both b_{xy} and b_{yx} are negative. Therefore r is negative.

(ii) Mean values of x and y is the point of intersection of the two lines of regression

$\therefore$ we get $(\bar{x}, \bar{y})$ by solving the equations

3x+12y=19 _____ (1)

3y + 9x=46 _____ (2)

$(1)\times 3, 9x + 36y = 57$

$9x + 3y = 46$

$33y = 11$

$$y = \frac{11}{33} = \frac{1}{3}$$

3x+12y=19

$3x+12.1/3=19 \Rightarrow 3x+4=19 \Rightarrow 3x=15$

$$x = \frac{15}{3} = 5$$

$$\therefore (\bar{x}, \bar{y}) = (5, 1/3)$$

$$\frac{\sigma_x^2}{\sigma_y^2} = r\frac{\sigma_y}{\sigma_x} = \frac{-3}{12} = \frac{-1}{4}$$

$$then, r^2 \frac{\sigma_y^2}{\sigma_x^2} = \frac{1}{16}, and, r^2 = \frac{1}{12}$$

$$\therefore \frac{\sigma_y^2}{\sigma_x^2} = \frac{1}{6}.\frac{1}{r^2} = \frac{12}{16} = \frac{3}{4}$$

$\therefore$ The ratio of the variance of x and variance of y is 3:4.

Example 8:

10 observations on price x and supply y the following data were obtained. $\sum x$ =130, $\sum y$ =220, $\sum x^2 = 2288, \sum y^2 = 5506, and, \sum xy = 3467$, obtain the line of regression of y on x and estimate the supply when the price is 16 units and also find the standard error of estimate.

Solution:

$$\bar{x} = \frac{130}{10} = 13, \sum x^2 = 2288$$

$$\sigma_x^2 = \frac{1}{n}\sum x^2 - \bar{x}^2 = \frac{2288}{10} - 169 = 59.8$$

$$\sigma_y^2 = \frac{1}{n}\sum y^2 - \bar{y}^2, \bar{y} = \frac{220}{10} = 22, \sum y^2 = 5506$$

$$= \frac{5506}{10} - 484 = 64.6$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum x_i y_i - \bar{x}\sum y_i - x_i \sum y_i + \bar{x}\bar{y}n$$

$$= 3467 - 13.220 - 22.130 + 13.22.10$$

$$= 607$$

$$r = \frac{S_{xy}}{n\sigma_x\sigma_y} = \frac{607}{10 \times 78 \times 8} = 0.99$$

$$y - 22 = 0.99\frac{8}{7.8}(x - 13)$$

$$y = 1.02x + 8.7$$

$$When prce, x = 16, y = 1.02 \times 16 + 8.7 = 25.02$$

$$r = 0.99$$

$$S.E = \frac{1 - r^2}{\sqrt{n}} = 0.006.$$

Example 9:

Calculate the coefficient of correlation from the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Also obtained the equations of the lines of regression and obtain on estimate of us which should correspond on the average to x = 6.2.

Solution:

The coefficient of correlation is $r = \frac{S_{xy}}{\sqrt{s_{xx}s_{yy}}}$ then

x_i	y_i	$x_i = x_i - \bar{x}$	$y_i = y_i - \bar{y}$	x_i^2	y_i^2	$x_i y_i$
1	9	-4	-3	16	9	12
2	8	-3	-4	9	16	12
3	10	-2	-2	4	4	4
4	12	-1	0	1	0	0
5	11	0	-1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	9
45	108			60	60	57

$$\bar{x} = \frac{45}{9} = 5, \bar{y} = \frac{108}{9} = 12$$

$$\sigma_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} = \frac{60}{9}, \sigma_y^2 = \frac{(y_i - \bar{y})^2}{n} = \frac{60}{9}$$

$$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{57}{\sqrt{60 \times 60}} = \frac{57}{60} = 0.95$$

The line of regression of y on x.

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$(y - 12) = 0.95 . \frac{\sqrt{60}}{3} . \frac{3}{\sqrt{60}}(x - 5) = 0.95(x - 5)$$

$$y - 0.95x = 12 - 4.75 = 7.25 - - - - - - - - (1)$$

The equation of line of regression of x on y is x-5 = 0.95 (y-12)

$$x - 0.95y = 5 - 11.40 = -6.1 - - - - - (2)$$

$$When, x = 6.2, substitutioninweget, y = 0.95 \times 6.2 + 7.25 = 13.14$$

Example 10:

Given n=10, $\sigma_x = 4.5, \sigma_y = 3.6$ and sum of product of deviation from the mean of x and y is find the correlation coefficient.

Solution:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

$$n = 10, \sigma_x = 4.5, \sigma_y = 3.6$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 64 given$$

$$\therefore r = \frac{64}{10 \times 4.5 \times 3.6} = 0.4$$

Example 11:

A chemical company wishing to study the effect of extraction fine on the efficiency of an extraction operation obtained in the data shown in the following table.

Extraction time x	27 45 41 19 35 39 19 49 15 31

Extraction efficiency y	57	64	80	46	62	72	52	77	57	68

Calculate r for the extraction times and extraction efficiencies. Assuming that the necessary assumptions can be met null hypothesis P =0.75 against the alternative hypothesis P>0.75 at 5% Level of significance.

Solutions:

In order to calculator we have

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

And test the null hypothesis H_o we have the test statistic

$$Z=\sqrt{n-3}z\frac{(1-p)}{(1+e)} = \frac{\sqrt{n-3}}{2}\ln\frac{1+r}{1-r}\left(\frac{1-p}{1+e}\right)$$

Since from alternative hypothesis H_1: P>0.75 a Right tailed test then Z. table value at 5% is = 1.645

27	57	-5	-6.5	25	42.25	32.5
45	64	13	0.5	169	0.25	6.5
41	80	9	16.5	81	272.25	148.5
19	46	-13	-17.5	169	306.25	227.5
35	62	3	-1.5	9	2.25	-4.5
39	72	7	8.5	49	72.25	59.5
19	52	-13	-11.5	169	132.25	149.5
49	77	17	13.5	289	182.25	229.5
15	57	-17	-6.5	289	42.25	110.5
31	68	-1	4.5	1	20.25	-4.5
320635		1250		1072.50	925	

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{925}{\sqrt{1250\times1072.5}} = 0.8$$

To list the significance, we set up null hypothesis H_o: P = 0.75

Alternative hypothesis H_i : P>7.5 (right tailed test)

Level of significance α =0.05

$$Z = \sqrt{n-3}z\frac{(1-\ell)}{(1+\ell)} = \frac{\sqrt{n-3}}{2}\ln\frac{1+r}{1-r}\left(\frac{1-\ell}{1+\ell}\right)$$

For r=0.8

$$\left(\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\right)=1.256(\ fromtables$$

$$\therefore Z=\left(\frac{1-0.75}{1+75}\right)\sqrt{10-3}\times 1.256=0.48$$

$$Z_{\alpha}=1.645at,5\%L.O.S$$

Conclusion: since Z calculated value 0.48 less z tabulated value 1.645 at 5% L.O.S. we accept null hypothesis $H_{o.}$

$\therefore H_o$: P = 0.75.

Example 12:

Assuming that the necessary assumption are met, construct 95% confidence interval for P when (i) = 0.72 and n=19, (ii) r = 0.57 and n=40

Solution: confidence interval for population correlation coefficient

$$p=z-\frac{z_{\alpha/2}}{\sqrt{n-3}}<\mu_2<z+\frac{z_{\alpha/2}}{\sqrt{n-3}}$$

$$z_{\alpha/2}=1.9, n=19$$

$$z,\ for, r=0.72,\ for 95\%=0.887$$

$$\therefore z-\frac{1.96}{\sqrt{16}}<\mu_2<z+\frac{1.96}{4}$$

$$0.887-0.49<\mu_2<0.887+0.49$$

$$0.397<\mu_2<1.377$$

R=0.57

$$\alpha=0.05, Z_{\alpha/2}=1.96\Rightarrow z=0.648\ fromtables$$

N=40, $\therefore 0.648-\frac{1.96}{\sqrt{37}}<\mu_2<0.648+1.96$

$$i.e, 0.328<\mu_2<0.968$$

Example 13:

The following show the improvement (gain in reading speed) of eight of student in a speed reading Program, and the number of weeks they have been in the program.

No. of weeks	3	5	2	8	6	9	3	4
Speed gain	86	118	49	193	164	232	73	109

a. Fit a straight line by the method of least squares.
b. Find a 90% confidence interval for β

Solution:

a) The straight line y=a+bx_____(1)
Normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

x_i	y_i	$x_i y_i$	x_i^2
3	86	258	9
5	118	590	25
2	49	98	4
8	193	1544	64
6	164	984	36
9	232	2088	81
3	73	219	9
4	109	436	16
40	10246	217	244

1024=8a+40b _____ (1)

6217 = 40a +244b _____ (2)

$(1)\times 5 \Rightarrow 5120 = 40a + 200b$

$6217 = 40a + 244b$

$1097 = 44b$

b = 24.93

a = 3.35

$\therefore$ y = 3.35 +24.93 x.

b. The straight line is y= 3.35 + 24.93x

confidence interval for $\beta = b \pm t_{\alpha/2} S_e \frac{1}{\sqrt{S_{xx}}}$

b = 24.93, n=8, n-2=6

$t_{\alpha/2}, \ for, bd. ffor 90\% confidence is, 1.94$

$$S_e^2 = \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{(n-2)}$$

$$\sum x^2 = 244, \sum x = 40, \sum y = 1024$$

$$\sum xy = 6217, \sum y^2 = 158900$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 244 - \frac{(40)^2}{8} = 27828$$

$$S_{xy} = \sum xy - \sum x.\frac{\sum y}{n} = 6217 - \frac{40.1024}{8} = 1097$$

$$S_e = \sqrt{\frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2}} = \sqrt{\frac{27828 - \frac{(1097)^2}{44}}{6}} = 9.3$$

$$\frac{t_{\alpha/2} S_e}{\sqrt{S_{xx}}} = \frac{1.94 \times 9.3}{\sqrt{44}} = 2.6$$

$$b \pm \frac{t_{\alpha/2} S_e}{\sqrt{S_{xx}}} = 22.33, and\, 27.53$$

$$\therefore 22.33 < \beta < 27.53$$

Example 14:

In the given table x is the tensile force applied to a steel specimen in thousands of pounds and y is the resulting elongation in thousands of an inch.

x	1	2	3	4	5	6
y	14	33	40	63	76	85

The equation of the least square line is y=1.12+14.49x

(i) construct 95% confidence interval for β, the elongation per thousand pounds of tensile stress

(ii) Find 95% limit of prediction for the elongation of a specimen with x= 3.5 thousand pounds.

Solution:

(i) given equation y=1.12+14.49x

Confidence interval for β is $b \pm t_{\alpha/2} S_e \frac{1}{\sqrt{S_{xx}}}$

B=14.49,n=6.

$$t_{\alpha/2} = t_{0.025}(ud.f) = 2.776$$

$$\sum_{i=1}^{n} x_i = 21, \sum_{i=1}^{n} y_i = 311, \sum_{i=1}^{n} x_i^2 = 91, \sum_{i=1}^{n} x_i y_i = 1342$$

$$\sum_{i=1}^{n} y_i^2 = 19855$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 91 - \frac{441}{6} = 17.5$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 19855 - \frac{(311)^2}{6} = 3734.8$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i . \sum_{i=1}^{n} \frac{y_i}{n} = 1342 - \frac{311 \times 21}{6} = 2535$$

$$S_e = \sqrt{\frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2}} = \sqrt{\frac{3734.8 - \frac{(253.5)^2}{17.5}}{4}} = 3.96$$

$$t_{\alpha/2} - S_e \frac{1}{\sqrt{S_{xx}}} = \frac{2.776 \times 3.96}{\sqrt{17.5}} = 2.63$$

∴ 95% confidence limits for β.

14.49±2.63 = (11.86,17.12).

b) 95% confidence limits for prediction for y when x = x_0=3.5

y=51.83 when x=3.5

$$y \pm t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$S_e = 3.96, t_{\alpha/2} = 2.776, x_0 = 3.5, \bar{x} = \frac{21}{6} 3.5$$

$$t_{\alpha/2} S_e \sqrt{\frac{1}{n} + 0} = 2.776 \times 3.96 \sqrt{\frac{1}{6}}$$

$$= 453251$$

$$y = (a + bx_0) \pm t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$
$$= 51.83 \pm 4.5325 = (47.2975, 56.3625)$$

Example 15:

The following table shows how many we a sample of six persons have worked at an auto mobile inspection station and the number of cars each one inspected between noon and 2pm on a given day

No. of weeks Employed	2	7	9	1	5	12
No. of cars Inspected	13	21	23	14	15	21

The line of best fit for the above data given by y=15.85+0.33x test the null hypothesis β=1.2

against the alternative hypothesis β<1.2 at the 0.05 level of significance.

Solution:

Null hypothesis $H_0 : \beta = 1.2$

Alternative hypothesis $H_1 : \beta < 1.2$

Level of significance $\alpha = 0.05$

This is left tail test.

$\therefore$ Test statistic to test the null hypothesis H_o is $t = \frac{b - \beta}{S_e} \sqrt{S_{xx}}$

$\beta = 1.2, b = 0.33, S_{xx} = 88, S_e = 4.57$

$y = 15.85 + 0.33x, given$

$then, t = \frac{(0.33 - 1.2)}{4.57} \sqrt{88} = -1.78$

$|t| = 1.78$

Conclusion: since H calculation value is less than t- tabulated value. Hence at 5% leve; of significance we accept our H_o. i.e., H_o; β=1.2.

Example 16:

Find the parabola of the form $y=a+bx+cx^2$ which fit most closely with the observations.

x	-3	-2	-1	0	1	2	3
y	4.63	2.11	0.67	0.09	0.63	2.15	4.58

Solution:

Normal equations are

$$\sum_{i=1}^{n} y_i = an + \sum_{i=1}^{n} x_i.b + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i. + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

x_i	y_i	$x_i y_i$				
-3	4.63	-13.89	9	-27	81	41.67
-2	2.11	-4.22	4	-8	16	8.44
-1	0.67	-0.67	1	-1	1	0.67
0	0.09	0	0	0	0	0
1	0.63	0.63	1	1	1	0.63
2	2.15	4.30	4	8	16	8.60
3	4.58	13.74	9	27	81	41.22

014.86-0.1128 0196101.23

14.86 = 7a+28c _____ (1)

-0.11 = 28b _____ (2)

101.23 = 28a+196c _____ (3)

59.44=28a+112c

41.79= 84c

∴c = 0.5

$$a = \frac{14.86 - 28c}{7} = 0.12$$

∴y = 0.12 – 0.004 x+ 0.5 x^2

Example 17:

Fit an equation of the form y=ab the following data.

x	2	3	4	5	6
y	144	172.8	207.4	248.8	298.5

Solution:

y=a b^x taking logarithm on both sides

logy = log a+ x log b

U= A+BX where U= log y, A= log a, B=log b the normal equations are

$$\sum U = nA + B\sum x, and, \sum XU = A\sum x + B\sum x^2$$

x=x	y	U = log y	X^2	xy
2	144	2.1584	4	4.3168
3	172.8	2.2375	9	6.7125
4	207.4	2.3168	16	9.2672
5	248.8	2.3959	25	11.9795
6	298.5	2.4749	36	14.8494
20		**11.5835**	**90**	**47.1254**

11.5835 = 5A + 20 B

47.1254 = 20A + 90 B

(1)×4 46.3340 = 20A + 80 B

47.1254 = 20 A + 90 B

0.7914 = 10 B

∴ B = 0.07914, A = $\frac{11.5835 - 20\,B}{5}$=2

a = 100, b=1.2

Hence y= 100 $(1.2)^x$

Example 18:

Find the curve of best fit of the type y= $a.e^{bx}$ to the following data by the method of least squares.

x	1	5	7	9	12
y	10	15	12	15	21

Solution:

The curve is y=$a.e^{bx}$

Taking log on both sides

logy = lag a + bxlog c

i.e., U = A+ B x where U = logy A= log a, B= b log c normal equations are

$$\sum U_i = nA + B\sum x$$
$$\sum x_i U_i = A\sum x_i + B\sum x_i^2$$

x_i	y_i	$U = \log y_i$	$x_i^2 = X_i^2$	$x_i y_i$
1	10	1.0000	1	1
5	15	1.1761	25	5.8805
7	12	1.0792	49	7.5544
9	15	1.1761	81	10.5849
12	21	1.3222	144	15.8664
34		5.7536	300	40.8862

n = 5, we have by using above normal equations

5.7536 = 5A + 34 B_____ (1)

40.8862 = 3A + 300 B _____ (2)

$(1)\times 34 \Rightarrow$ 195.6224=170A+1156B

$(2)\times 5 \Rightarrow$ 204.4310=170A+1500B

8.8086= 344B

$$B = \frac{8.8086}{344} = 0.0256$$

$$A = \frac{5.7536 - 34(0.0256)}{5} = 0.9766$$

A=9.4754, b=0.059, y=9.4754 $e^{0.059}$

Example 19:

Fit the model $y= ax^b$ to the following data.

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.10	6.80	7.50

Solution: $y = ax^b$

Taking logarithm on both sides we set

Log y = log a + b log x

U = A + Bv where U = log y, A=log a v=log x

∴normal equations are

$$\sum U_i = nA + b\sum V_i$$
$$\sum U_i V_i = A\sum V_i + b\sum v_i^2$$

x	y	V = log x	U = log y	UV	V^2
1	2.98	0	0.4742	0	0
2	4.26	0.3010	0.6294	0.1894	0.0906
3	5.21	0.4771	0.7168	0.3420	0.2276
4	6.10	0.6.21	0.7853	0.4728	0.3625
5	6.80	0.6990	0.8325	0.5819	0.4886
6	7.50	0.7782	0.8751	0.6810	0.6056
		2.8574	4.3133	2.2671	1.7749

∴The normal equations are

4.3133 = 6A +2.8574 b

2.2671 = 2.8574A + 1.7749 b

By solving above two equations we set

B =0.5142, A = 0.4740, log a = 0.4740, a = anti log (0.4740) = 2.978

∴$y = ax^b = 2.978 \times 0.5142$

Example 20:

Find the least squares regression equation of x_1 on x_2 and x_3 from the following data.

x_1	3	5	6	8	12	14
x_2	16	10	7	4	3	2
x_3	90	72	54	42	30	12

Solution:

Let $x_1 = a_0 + a_1x_2 + a_2x_3$

Changing the origin $u = x_2 - 7$ and $v = x_3 - 50$

Let $x_1 = a + bu + cv$

The normal equations are

$$\sum_{i=1}^{n} x_{1i} = na + b\sum_{i=1}^{n} u_i + c\sum_{i=1}^{n} V_i$$

$$\sum_{i=1}^{n} x_{1i}u_i = a\sum_{i=1}^{n} u_i + b\sum_{i=1}^{n} u_i^2 + c\sum_{i=1}^{n} u_i v_i$$

$$\sum_{i=1}^{n} x_{1i}u_i = a\sum_{i=1}^{n} u_i + b\sum_{i=1}^{n} u_i v_i + c\sum_{i=1}^{n} v_i^2$$

x_1	x_2	x_3	u_i	v_i	X_{1i} U_i	X_{1i} V_i	U_i V_i	u_i^2	v_i^2
3	16	90	9	40	27	120	360	81	1600
5	10	72	3	22	15	110	66	9	484
6	7	54	0	4	0	24	0	0	16
8	4	42	-3	-8	-24	-64	24	9	64
12	3	30	-4	-20	-48	-240	80	16	400
14	2	12	-5	-38	-70	-532	190	25	1444
48		0	0	-100	-100	-582	720	140	4008

n= 6 48=6a +0+0 _____ (1)

$\therefore$ a = 8

-100 = 140 b +720 c _____ (2)

-582 = 720 b +4008 c _____ (3)

$(2)\times 36$ -3600=5040b+25920c

-4074=5040b+28030c

474=-2136c

$\therefore c = -0.22$

$$b = \frac{100 - 720(-0.22)}{140} = 0.417$$

$$y = 8 + 0.4174 - 0.22v$$
$$= 8 + 0.417(x_2 - 7) - 0.22(x_3 - 50)$$
$$= 8 + 0.417x_2 - 2.919 - 0.22x_2 + 11$$
$$= 16.1 + 0.417x_2 - 0.22x_3$$

Example 21:

The following are data on the number of twists required to break a certain kind of forged alloy far and the percentages of two alloying elements present in the metal with following elements present in the metal with the following data.

$$\sum x_{1i} = 40, \sum x_{2i} = 200, \sum x_{1i}^2 = 120, \sum x_{1i}x_{2i} = 500, \sum x_{2i}^2 = 3{,}000, \sum y_i = 723, \sum x_{1i}y_i = 1{,}963$$

$\sum x_i y_i = 8{,}210$

With normal equation

$723 = 16a_0 + 40a_1 + 200a_2$
$1{,}963 = 40a_0 + 120a_1 + 500a_2$
$8210 = 200a_0 + 500a_1 + 3000a_2$

Use the matrix expressions to data muil the least squares estimates of the multiple regression coefficient.

Solution:

Since we are given $\sum x_{1i}^2 = 120, \sum x_{1i}x_{2i} = 500, \sum x_{2i} = 200, and,$

and n=16 are substituted in to the expression of X^1x given by

$$X^1X = \begin{bmatrix} n & \Sigma x_{1i} & \Sigma x_{2i} \\ \Sigma x_{1i} & \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{2i} & \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} = \begin{bmatrix} 16 & 40 & 200 \\ 40 & 120 & 500 \\ 200 & 500 & 3000 \end{bmatrix}$$

Then the inverse of this matrix can be obtained by any one of a number of techniques and using the one based on cofactors, we find that

$$(XX)^{-1} = \frac{1}{160000}\begin{bmatrix} 110000 & -20000 & -4000 \\ -20000 & 8000 & 0 \\ -4000 & 0 & 320 \end{bmatrix}$$

Where 160,000 is the value of $|X^1X|$, the determinant of X^1X . Substitute

$\Sigma y_i = 723, \Sigma x_{1i}y_i = 1963\ and\ \Sigma x_{2i}y_i = 8210$into the expression of X^1y, we then get

$$X^1y = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix} = \begin{bmatrix} 723 \\ 1963 \\ 8210 \end{bmatrix}$$

And finally

$$b = (X^1X)^{-1}X^1y = \frac{1}{160000}\begin{bmatrix} 110000 & -20000 & -4000 \\ -20000 & 8000 & 0 \\ -4000 & 0 & 320 \end{bmatrix}\begin{bmatrix} 723 \\ 1963 \\ 8210 \end{bmatrix}$$

$$= \frac{1}{160000}\begin{bmatrix} 7430000 \\ 1244000 \\ -264800 \end{bmatrix} = \begin{bmatrix} 46.4375 \\ 7.7750 \\ -1.6550 \end{bmatrix}$$

Example 22:

Find 95% confidence for β_1 given β_1=1.61591, C_{11} =0.00274378, $\hat{\sigma}^2$ = 10.6239,

$t_{0.025,22}$ = 2.074.

Solution:

The confidence in multiple regressions is given by

$$\hat{\beta}_j \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

Given $\hat{\beta}_1 = 1.61591, C_{11} = 0.00274378, \hat{\sigma}^2 = 10.6239, t_{0.025,22} = 2.074$

$then, 1.26181 \le \beta_1 \le 1.97001$

4.5 Exercise

1. Fit the curve y= ae^{bx} to the following data

x	0.0	0.5	1.0	1.5	2.0	2.5
y	0.10	0.45	2.15	9.15	40.35	180.75

Ans: $y = 0.1019e^{2.9963x}$

2. Fit y= $a.b^x$ by the method of least squares to the following data

x	0	1	2	3	4	5	6	7
y	10	21	35	59	92	200	400	610

Ans: $y = 10.499(1.7959)^x$

3. fit a parabola for the data

x	1	2	3	4	5
y	1090	1220	1390	1625	1915

Ans: $y = 1024 + 40.5x + 27.5x^2$

Fit a straight line y= a+ bx and along a parabola to the following set of observations calculate the sum of squares of residuals in each case and test which curve is more suitable to the data.

x	0	1	2	3	4
y	1	5	10	22	38

(Ans: y= 9.1x- 3, y= 1.42+.26x+2.21x^2, parabola is the best curve)

5. The following sample data were collected to determine the relationship between two processing variables and the current gain of a certain kind of transistor, fit the least squares regression equation of y on x_1 and x_2. And use the matrix notation also

x_1	1.5	2.5	0.5	1.2	2.6	0.3	2.4	2.0	0.7	1.6
x_2	66	87	69	141	93	105	111	78	66	123
y	5.3	7.8	7.4	9.8	10.8	9.1	8.1	7.2	6.5	12.6

(Ans: $y = 2.3+0.23x_1 + 0.06x_2$)

6. Calculate the correlation coefficient for the heights of fathers and their sons.

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

(Ans: = 0.603)

7.

No. of weeks employed x	2	7	9	1	5	12
No. of weeks employed y	13	21	23	14	15	21

With reference to the above data the straight line of best fit in y = 15.85 + 0.33 x (1) find a 95% confidence interval for the average number of cars inspected in the given period of time by a person who has been working at the inspection station for 8 weeks (ii) 95% limits of prediction for the number of cars that will be inspected in the given period of time by a person who has worked at the inspection station for 8 weeks.

8. the two regression lines are having their mean, standard deviations 31.6,38 and 3.72 and 6.31and r =-0.36.

Find the two regression lines.

4.6 Summary

In this unit an attempt is made to explain the concepts of simple linear regression, correlation, and multiple regression models, estimation, test, interval procedure associated with them along with both theory and practical. A number of example are workout and a good number of exercises are also given.

4.7 Technical Terms:

Simple linear regression

Correlation

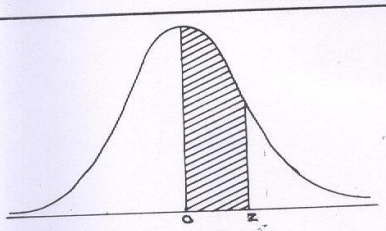
Multiple linear regressions

Least squares procedure

Matrix approach to least squares procedures.

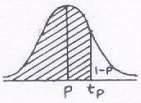
Interval estimation.

41



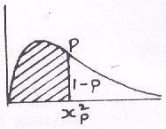
Areas under the standard Normal Curve from 0 to z.

z	0	1	2	3	4	5	6	7	8	9
0·0	·0000	·0040	·0080	·0120	·0160	·0199	·0239	·0279	·0319	·0359
0·1	·0398	·0438	·0478	·0517	·0557	·0596	·0636	·0675	·0714	·0754
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	·2258	·2291	·2324	·2357	·2389	·2422	·2454	·2486	·2518	·2549
0·7	·2580	·2612	·2642	·2673	·2704	·2734	·2764	·2794	·2823	·2852
0·8	·2881	·2910	·2939	·2967	·2996	·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	·4098	·4115	·4131	·4147	·4162	·4177
1.4	·4912	·4207	·4222	·4236	·4251	·4256	·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	·4545
1.7	·4554	·4564	·4573	·4582	·4591	·4599	·4608	·4616	·4625	·4633
1.8	·4641	·4649	·4656	·4664	·4671	·4678	·4686	·4693	·4699	·4706
1.9	·4713	·4719	·4726	·4732	·4738	·4744	·4750	·4756	·4761	·4767
2.0	·4772	·4778	·4783	·4788	·4793	·4798	·4803	·4808	·4812	·4817
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	·4916
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	·4963	·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	·4987	·4987	·4987	·4988	·4988	·4989	·4989	·4989	·4990	·4990



Percentile Values (t_p) for Student's t Distribution with v Degrees of Freedom

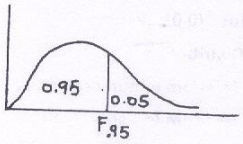
v	$t_{.55}$	$t_{.60}$	$t_{.70}$	$t_{.75}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	·158	·325	·727	1·000	1·376	3·08	6·31	12·71	31·82	63·66
2	·142	·289	·617	·816	1·061	1·89	2·92	4·30	6·96	9·92
3	·137	·277	·584	·765	·978	1·64	2·35	3·18	4·54	5·84
4	·134	·271	·569	·741	·941	1·53	2·13	2·78	3·75	4·60
5	·132	·267	·559	·727	·920	1·48	2·02	2·57	3·36	4·03
6	·131	·265	·553	·718	·906	1·44	1·94	2·45	3·14	3·71
7	·130	·263	·549	·711	·896	1·42	1·90	2·36	3·00	3·50
8	·130	·262	·546	·706	·889	1·40	1·86	2·31	2·90	3·36
9	·129	·261	·543	·703	·883	1·38	1·83	2·26	2·82	3·25
10	·129	·260	·542	·700	·879	1·37	1·81	2·23	2·76	3·17
11	·129	·260	·540	·697	·876	1·36	1·80	2·20	2·72	3·11
12	·128	·259	·539	·695	·873	1·36	1·78	2·18	2·68	3·06
13	·128	·259	·538	·694	·870	1·35	1·77	2·16	2·65	3·01
14	·128	·258	·537	·692	·868	1·34	1·76	2·14	2·62	2·98
15	·128	·258	·536	·691	·866	1·34	1·75	2·13	2·60	2·95
16	·128	·258	·535	·690	·865	1·34	1·75	2·12	2·58	2·92
17	·128	·257	·534	·689	·863	1·33	1·74	2·11	2·57	2·90
18	·127	·257	·534	·688	·862	1·33	1·73	2·10	2·55	2·88
19	·127	·257	·533	·688	·861	1·33	1·73	2·09	2·54	2·86
20	·127	·257	·533	·687	·860	1·32	1·72	2·09	2·53	2·84
21	·127	·257	·532	·686	·859	1·32	1·72	2·08	2·52	2·83
22	·127	·256	·532	·686	·858	1·32	1·72	2·07	2·51	2·82
23	·127	·256	·532	·685	·858	1·32	1·71	2·07	2·50	2·81
24	·127	·256	·531	·685	·857	1·32	1·71	2·06	2·49	2·80
25	·127	·256	·531	·684	·856	1·32	1·71	2·06	2·48	2·79
26	·127	·256	·531	·684	·856	1·32	1·71	2·06	2·48	2·78
27	·127	·256	·531	·684	·855	1·31	1·70	2·05	2·47	2·77
28	·127	·256	·530	·683	·855	1·31	1·70	2·05	2·47	2·76
29	·127	·256	·530	·683	·854	1·31	1·70	2·04	2·46	2·76
30	·127	·256	·530	·683	·854	1·31	1·70	2·04	2·46	2·75



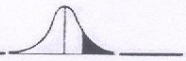
Percentile Values (x^2_p) for the Chi-Square Distribution with v Degree of Freedom

v	$x^2_{.005}$	$x^2_{.01}$	$x^2_{.025}$	$x^2_{.05}$	$x^2_{.10}$	$x^2_{.25}$	$x^2_{.50}$	$x^2_{.75}$	$x^2_{.90}$	$x^2_{.95}$	$x^2_{.975}$	$x^2_{.99}$	$x^2_{.995}$	$x^2_{.999}$
1	·0000	·0002	·0010	·0039	·0158	·102	·455	1·32	2·71	3·84	5·02	6·63	7·88	10·8
2	·0100	·0201	·0506	·103	·211	·575	1·39	2·77	4·61	5·99	7·38	9·21	10·6	13·8
3	0·717	·115	·216	·352	·584	1·21	2·37	4·11	6·25	7·81	9·35	11·3	12·8	16·3
4	·207	·297	·484	·711	1·06	1·92	3·36	5·39	7·78	9·49	11·1	13·3	14·9	18·5
5	·412	·554	·831	1·15	1·61	2·67	4·35	6·63	9·24	11·1	12·8	15·1	16·7	20·5
6	·676	·872	1·24	1·64	2·20	3·45	5·35	7·84	10·6	12·6	14·4	16·8	18·5	22·5
7	·989	1·24	1·69	2·17	2·83	4·25	6·35	9·04	12·0	14·1	16·0	18·5	20·3	24·3
8	1·34	1·65	2·18	2·73	3·49	5·07	7·34	10·2	13·4	15·5	17·5	20·1	22·0	26·1
9	1·73	2·09	2·70	3·33	4·17	5·90	8·34	11·4	14·7	16·9	19·0	21·7	23·6	27·9
10	2·16	2·56	3·25	3·94	4·87	6·74	9·34	12·5	16·0	18·3	20·5	23·2	23·2	29·6
11	2·60	3·05	3·82	4·57	5·58	7·58	10·3	13·7	17·3	19·7	21·9	24·7	26·8	31·3
12	3·07	3·57	4·40	5·23	6·30	8·44	11·3	14·8	18·5	21·0	23·3	26·2	28·3	32·9
13	3·57	4·11	5·01	5·89	7·04	9·30	12·3	16·0	19·8	22·4	24·7	27·7	29·8	34·5
14	4·07	4·66	5·63	6·57	7·79	10·2	13·3	17·1	21·1	23·7	26·1	29·1	31·3	36·1
15	4·60	5·23	6·26	7·26	8·55	11·0	14·3	18·2	22·3	25·0	27·5	30·6	32·8	37·7
16	5·14	5·81	6·91	7·96	9·31	11·9	15·3	19·4	23·5	26·3	28·8	32·0	34·3	39·3
17	5·70	6·41	7·56	8·67	10·1	12·8	16·3	20·5	24·8	27·6	30·2	33·4	35·7	40·8
18	6·26	7·01	8·23	9·39	10·9	13·7	17·3	21·6	26·0	28·9	31·5	34·8	37·2	42·3
19	6·84	7·63	8·91	10·1	11·7	14·6	18·3	22·7	27·2	30·1	32·9	36·2	38·6	43·8
20	7·43	8·26	9·59	10·9	12·4	15·5	19·3	23·8	28·4	31·4	34·2	37·6	40·0	45·3
21	8·03	8·90	10·3	11·6	13·2	16·3	20·3	24·9	29·6	32·7	35·5	38·9	41·4	46·8
22	8·64	9·54	11·0	12·3	14·0	17·2	21·3	26·0	30·8	33·9	36·8	40·3	42·8	48·3
23	9·26	10·2	11·7	13·1	14·8	18·1	22·3	27·1	32·0	35·2	38·1	41·6	44·2	49·7
24	9·89	10·9	12·4	13·8	15·7	19·0	23·3	28·2	33·2	36·4	39·4	43·0	45·6	51·2
25	10·5	11·5	13·1	14·6	16·5	19·9	24·3	29·3	34·4	37·7	40·6	44·3	46·9	52·6
26	11·2	12·2	13·8	15·4	17·3	20·8	25·3	30·4	35·6	38·9	41·9	45·6	48·3	54·1
27	11·8	12·9	14·6	16·2	18·1	21·7	26·3	31·5	36·7	40·1	43·2	47·0	49·6	55·5
28	12·5	13·6	15·3	16·9	18·9	22·7	27·3	32·6	37·9	41·3	44·5	48·3	51·0	56·9
29	13·1	14·3	16·0	17·7	19·8	23·6	28·3	33·7	39·1	42·6	45·7	49·6	52·3	58·3
30	13·8	15·0	16·8	18·5	20·6	24·5	29·3	34·8	40·3	43·8	47·0	50·9	53·7	59·7
40	20·7	22·2	24·4	26·6	29·1	33·7	39·3	45·6	51·8	55·8	59·3	63·7	66·8	73·4
50	28·0	29·7	32·4	34·8	37·7	42·9	49·3	56·3	63·2	67·5	71·4	76·2	79·5	86·7
60	35·5	37·5	40·5	43·2	46·5	52·3	59·3	67·0	74·4	79·1	83·3	88·4	92·0	99·6
70	43·3	45·4	48·8	51·7	55·3	61·7	69·3	77·6	85·5	90·5	95·0	100	104	112
80	51·2	53·5	57·2	60·4	64·3	71·1	79·3	88·1	96·6	120	107	112	116	125
90	59·2	61·8	65·6	69·1	73·3	80·6	89·3	98·6	108	113	118	128	128	137
100	67·3	70·1	74·2	77·9	82·4	90·1	99·9	109	118	124	130	136	140	149

9[illegible] centile Values (0·05 Levels),
$F_{.95}$ [illegible] the F Distribution
v_1 degrees of freedom in numerator
v_2 degrees of freedom in denominator.



v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18·5	19·0	19·2	19·2	19·3	19·3	19·4	19·4	19·4	19·4	19·4	19·4	19·4	19·5	19·5	19·5	19·5	19·5	19·5
3	10·1	9·55	9·28	9·12	9·01	8·94	8·89	8·85	8·81	8·79	8·74	8·70	8·66	8·64	8·62	8·59	8·57	8·55	8·53
4	7·71	6·74	6·59	6·39	6·26	6·16	6·09	6·04	6·00	5·96	5·91	5·86	5·80	5·77	5·75	5·72	5·69	5·66	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·88	4·82	4·77	4·74	4·68	4·62	4·56	4·53	4·50	4·46	4·43	4·40	4·37
6	5·99	5·14	4·76	4·53	4·39	4·28	4·21	4·15	4·10	4·06	4·00	3·94	3·87	3·84	3·81	3·77	3·74	3·70	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·79	3·73	3·68	3·64	3·57	3·51	3·44	3·41	3·38	3·34	3·30	3·27	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·50	3·44	3·39	3·35	3·28	3·22	3·15	3·12	3·08	3·04	3·01	2·97	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·29	3·23	3·18	3·14	3·07	3·01	2·94	2·90	2·86	2·83	2·79	2·75	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·14	3·07	3·02	2·98	2·91	2·85	2·77	2·74	2·70	2·66	2·62	2·58	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	3·01	2·95	2·90	2·85	2·79	2·72	2·65	2·61	2·57	2·53	2·49	2·45	2·40
12	4·75	3·89	2·49	3·26	3·11	3·00	2·91	2·85	2·80	2·75	2·69	2·62	2·54	2·51	2·47	2·43	2·38	2·34	2·30
13	4·67	3·81	3·41	3·18	3·03	2·92	2·83	2·77	2·71	2·67	2·60	2·53	2·46	2·42	2·38	2·34	2·30	2·25	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·76	2·70	2·65	2·60	2·53	2·46	2·39	2·35	2·31	2·27	2·22	2·18	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·71	2·64	2·59	2·54	2·48	2·40	2·33	2·29	2·25	2·20	2·16	2·11	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·66	2·59	2·54	2·49	2·52	2·35	2·28	2·24	2·19	2·15	2·11	2·06	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·61	2·55	2·49	2·45	2·38	2·31	2·23	2·19	2·15	2·10	2·06	2·01	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·58	2·51	2·46	2·41	2·34	2·27	2·19	2·15	2·11	2·06	2·02	1·97	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·54	2·48	2·42	2·38	2·31	2·23	2·16	2·11	2·07	2·03	1·98	1·93	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·51	2·45	2·39	2·35	2·28	2·20	2·12	2·08	2·04	1·99	1·95	1·90	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·49	2·42	2·37	2·32	2·25	2·18	2·10	2·05	2·01	1·96	1·92	1·87	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·46	2·40	2·34	2·30	2·23	2·15	2·07	2·03	1·98	1·94	1·89	1·84	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·44	2·37	2·32	2·27	2·20	2·13	2·05	2·01	1·96	1·91	1·86	1·81	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·42	2·36	2·30	2·25	2·18	2·11	2·03	1·98	1·94	1·89	1·84	1·79	1·73
25	4·24	3·39	2·99	2·76	2·60	2·49	2·40	2·34	2·28	2·24	2·16	2·09	2·01	1·96	1·92	1·87	1·82	1·77	1·71
26	4·23	3·37	2·98	2·74	2·59	2·47	2·39	2·32	2·27	2·22	2·15	2·07	1·99	1·95	1·90	1·85	1·80	1·75	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·37	2·31	2·25	2·20	2·13	2·06	1·97	1·93	1·88	1·84	1·79	1·73	1·67
28	4·20	3·34	2·95	2·71	2·56	2·45	2·36	2·29	2·24	2·19	2·12	2·04	1·96	1·91	1·87	1·82	1·77	1·71	1·65
29	4·18	3·33	2·93	2·70	2·55	2·43	2·35	2·28	2·22	2·18	2·10	2·03	1·94	1·90	1·85	1·81	1·75	1·70	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·33	2·27	2·21	2·16	2·09	2·01	1·93	1·89	1·84	1·79	1·74	1·68	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·25	2·18	2·12	2·08	2·00	1·92	1·84	1·79	1·74	1·69	1·64	1·58	1·51
60	4·00	3·15	2·76	2·53	2·37	2·25	2·17	2·10	2·04	1·99	1·92	1·84	1·75	1·70	1·65	1·59	1·53	1·47	1·39
120	3·92	3·07	2·68	2·45	2·29	2·18	2·09	2·02	1·96	1·91	1·83	1·75	1·66	1·61	1·55	1·50	1·43	1·35	1·25
∞	3·84	3·00	2·60	2·37	2·21	2·10	2·01	1·94	1·88	1·83	1·75	1·67	1·57	1·52	1·46	1·39	1·32	1·22	1·00



Values of $Z = \frac{1}{2} \ln \frac{1 + r}{1 - r}$

r	0·00	0·01	0·02	0·03	0·04	0·05	0·06	0·07	0·08	0·09
0·0	0·000	0·010	0·020	0·030	0·040	0·050	0·060	0·070	0·080	0·090
0·1	0·100	0·110	0·121	0·131	0·141	0·151	0·161	0·172	0·182	0·192
0·2	0·203	0·213	0·224	0·234	0·245	0·255	0·266	0·277	0·288	0·299
0·3	0·310	0·321	0·332	0·343	0·354	0·365	0·377	0·388	0·400	0·412
0·4	0·424	0·436	0·448	0·460	0·472	0·485	0·497	0·510	0·523	0·536
0·5	0·549	0·563	0·576	0·590	0·604	0·618	0·633	0·648	0·662	0·678
0·6	0·693	0·709	0·725	0·741	0·758	0·775	0·793	0·811	0·829	0·848
0·7	0·867	0·887	0·908	0·929	0·950	0·973	0·996	1·020	1·045	1·071
0·8	1·099	1·127	1·157	1·188	1·221	1·256	1·293	1·333	1·376	1·422
0·9	1·472	1·528	1·589	1·658	1·738	1·832	1·946	2·092	2·298	2·647