KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University Under Section3 of UGC Act 1956)
COIMBATORE-641 021
FACULTY OF ENGINEERING
DEPARTMENT OF SCIENCE AND HUMANITIES
Syllabus

16BTBT301, 16BTCE301, 16BEBME401

PROBABILITY AND STATISTICS 3 2 0 4

OBJECTIVES:

1. To introduce the concept of probability and Sampling techniques.
2. To understand the fundamentals of Experimental Designs and Quality Control.

INTENDED OUTCOME:

1. The students would be exposed to statistical methods designed to contribute to the process of making scientific judgments in the face of uncertainty and variation.

UNIT I PROBABILITY (11)
Probability – Definition – Law - conditional probability-Bayes theorem- Probability mass function - Probability density functions.

UNIT II RANDOM VARIABLES (13)
Introduction to one dimensional random variables – Discrete – Continuous - Joint distributions – Marginal and conditional distributions – Covariance – Correlation and Regression.

UNIT III TESTING OF HYPOTHESIS (12)
Sampling distributions – Testing of hypothesis for mean, variance, proportions and differences using Normal, t, Chi-square and F distributions – Tests for independence of attributes and Goodness of fit.

UNIT IV DESIGN OF EXPERIMENTS (12)
Analysis of variance – one way classification – CRD – Two-way classification – RBD – Latin square.

UNIT V RELIABILITY AND QUALITY CONTROL (12)
Concepts of reliability – hazard functions – Reliability of series and parallel systems – control charts for measurement ($\bar{X}$ *and* R *charts*) - Control charts for attributes (p, c and np charts).

Total: 60

TEXT BOOKS:

S. NO.	AUTHOR(S) NAME	TITLE OF THE BOOK	PUBLISHER	YEAR OF PUBLICATION
1	R.A.Johnson and C.B.Gupta	Miller and Freund's Probability and Statistics for Engineers	Pearson Education Asia, New Delhi.	2007
2	S.C.Gupta and V.K.Kapoor	Fundamentals of Mathematical Statistics	Sultan Chand & Sons, New Delhi	2014

REFERENCES:

S. NO.	AUTHOR(S) NAME	TITLE OF THE BOOK	PUBLISHER	YEAR OF PUBLICATION
1	P.S.S.SundarRao and J.Richard	Introduction to Biostatistics and Research Methods	Prentice Hall of India, New Delhi.	2012
2	S.C.Gupta and V.K.Kapoor	Fundamentals of Applied Statistics	Sultan Chand & Sons, New Delhi	2007

WEBSITES:

1. www.cut-theknot.org/probability.shtml
2. www.mathcentre.ac.uk
3. www.mathworld. Wolfram.com

Staff -incharge HoD

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University Under Section3 of UGC Act 1956)
COIMBATORE-641 021
FACULTY OF ENGINEERING
DEPARTMENT OF SCIENCE AND HUMANITIES

LECTURE PLAN

Subject : PROBABILITY AND STATISTICS
Code : 16BEBME401

Unit No.	List of Topics	No. of Hours
UNIT I	**PROBABILITY**	
	Introduction-Basics of Probability-Random experiments, trial, types of events-Problems	1
	Probability- Definition –Axioms of Probability and law	1
	Axioms of Probability and law	1
	Concept of Conditional Probability	1
	Conditional Probability-problems	1
	Baye's theorem	1
	Problems based on Baye's theorem	1
	Introduction to random variables-Examples	1
	Probability mass function	1
	Problems based on Probability mass function	1
	Probability density function	1
	Problems based on Probability density function	1
	TOTAL	**12**
UNIT – II	**RANDOM VARIABLES**	
	Introduction to one dimensional random variables	1
	Problems in one dimensional random variables	1
	Discrete random variable-Problems	1
	Continuous random variable-Problems	1
	Introduction to Joint distributions and problems	1
	Discrete and Continuousrandom variable Problems	1
	Problems based on Joint distributions	1
	Problems based on Marginal distributions and Conditional distributions	1
	Concepts of Covariance	1
	Covariance problems	1
	Introduction to Correlation and Regression	1
	Correlation and Regression-Problems	1
	Correlation and Regression-Problems	1
	TOTAL	**13**
UNIT – III	**TESTING OF HYPOTHESIS**	
	Introduction to sampling distributions	1
	Concepts of testing of hypothesis using normal distribution	1
	Problems based on testing of hypothesis using normal distribution	1
	Concepts of testing of hypothesis using t distribution	1
	Concepts of based on testing of hypothesis using Chi-square distributions and Goodness of fit	1
	Problems based on testing of hypothesis using Chi-square distributions and Goodness of fit	1

	Testing of hypothesis using Chi-square distributions, independence of attributes	1
	Problems based on testing of hypothesis using Chi-square distributions, independence of attributes	1
	Problems based on testing of hypothesis using Chi-square distributions, independence of attributes	1
	Concept of testing of hypothesis using F-distributions	1
	Axioms of Probability and law	
	Axioms of Probability and law	1
	TOTAL	**12**
UNIT – IV	**DESIGN OF EXPERIMENTS**	
	Introduction to design of experiments	1
	Analysis of variance	1
	Concept of one way classification	1
	Completely Randomized Design	1
	Problems - Completely Randomized Design	1
	Problems - Completely Randomized Design	1
	Concept of Two-way classification	1
	Problems based on Randomized Block Design	1
	Problems based on Randomized Block	1
	Latin square Design	1
	Problems based on Latin square Design	1
	Problems based on Latin square Design	1
	TOTAL	**12**
UNIT – V	**RELIABILITY AND QUALITY CONTROL**	
	Introduction to reliability and quality control	1
	Concepts of reliability	1
	Hazard functions	1
	Problems based on Hazard functions	1
	Problems based on Reliability of series and parallel systems	1
	Problems based on Reliability of series and parallel systems	1
	Control charts for measurement ($\bar{X}$ *and* R *charts*)	1
	Control charts for measurement ($\bar{X}$ *and* R *charts*)	1
	Problems based on Control charts for attributes (p,c and np charts).	1
	Problems based on Control charts for attributes (p,c and np charts).	1
	Problems based on Control charts for attributes (p,c and np charts).	1
	TOTAL	**11**
	TOTAL NO OF HOURS	**60**

staff-incharge HoD

UNIT-I

PROBABILITY

Introduction:

The word 'Probability or change' is very frequency used in day-to-day conversation. The Statistician I.J. Good, suggests in his "kinds of Probability" that "the theory of Probability is much older than the human species.

The concept and applications of probability, which is a formal term of the popular word "Change" while the ultimate objective is to facilitate calculation of probabilities in business and managerial, science and technology etc., the specific objectives are to understand the following terminology.

Random Experiment: The term experiment refers to describe, which can be repeated under some given conditions. The experiment whose result (outcomes) depends on change is called Random Experiment.

Example:

1. Tossing of a coin is a random experiment.
2. Throwing a die is a random experiment.
3. Calculation of he mean arterial blood pressure of a person under ideal environmental conditions, by using the formula, Blood pressure $= \frac{Systoloic\ pressure}{Diastolic\ pressure}\ mm/Hg$ is a random experiment.

Sample Space:

The totality of all possible outcomes of a random experiment is called a sample space and it is denoted by s and a possible outcome are element.

The no. of the coins in a sample space denoted by n(s).

Example:

Tossing a coin n(s)=2={H,T}

Event:

The output or result of a random experiment is called an event or result or outcome.

Example:

1. In tossing of a coin, getting head or tail is an event.
2. In throwing a die getting 1 or 2 or 3 or 4 or 5 or 6 is an event.

Events are generally denoted by capital letters A, B, C etc. The events can be of two types. One is simple event and the other is compound event

Favorable event:

The no. of events favorable to an event in a trail is the no.of outcomes which entire the happening of the event.

Mutually Exclusive Events:

Two or more events are said to be mutually exclusive events if the occurrence of one event precludes (excludes or prevents) the occurrence of others, i.e., both cannot happen simultaneously in a single trail.

Example:

1. In tossing of a coin, the events head and tail are mutually exclusive.
2. In throwing a die, all the six faces are mutually exclusive.

Equally Likely Events: Two or more events are said to be equally likely, if there is no reason to expect any one case (or any event) in preference to others. i.e., every outcome of the experiment has equal possibility of occurrence. These are equally likely events.

Exhaustive Number of Cases or Events: The total number of possible outcomes in an experiment is called exhaustive number of cases or events.

Dependent event:

Two events are said to be dependent if the occurance or non occurance of a event in any trail affect the occurance of the other event in other trail.

Classical Definition of Probability: Suppose that an event 'A' can happen in 'm' ways and fails to happen (or non-happen) in 'n' ways, all these 'm+n' ways are supposed equally likely. Then the probability of occurrence (or happening) of the event called its success is denoted by 'P(A)' or simply 'p' and is defined as $P(A)=\frac{m}{m+n}\ ...(1)$ and the probability of non-occurrence (or non-happening) of the event called its failure is denoted by $P(\overline{E})$ or simply 'q' and is defined as. $P(\overline{A})=\frac{n}{m+n}...(2)$

From (1) and (2) we observe that the probability of an event can be defined as

$$P(event)=\frac{The\ number\ of\ favourable\ cases\ for\ the\ event}{Total\ number\ of\ possible\ cases}$$

Definition:

Let S be the sample space and A be the event associated with a random experiment. Let n(S) and n(A) be the no .of elements of S & A. Then the probability of the event A occurring denoted as P(A) is defined by

$$P(event)=\frac{The\ number\ of\ favourable\ cases\ for\ the\ event}{Total\ number\ of\ possible\ cases}=\frac{n(A)}{n(S)}$$

Note:

It follows that, $P(A)+P(\overline{A})=1 or\ p+q=1.$

This implies that p=1-q or q=1-p.

Hence $0 \le P(A) \le 1$.

Axiomatic Definition of Probability: Let S be the sample space and A be an event associated with a random experiment. Then the probability of the event A, denoted by P(A), is defined as a real number satisfying he following axioms.

(i) $0 \le P(A) \le 1$

(ii) P(S)=1

(iii) If A and B are mutually exclusive events, $P(A \cup B)=P(A)+P(B)$

(iv) If $A_1, A_{2,...,} A_{n,...}$ are a set of mutually exclusive events, $P(A_1 \cup A_2 \cup ... \cup A_{n,...})=P(A_1)+P(A_2)+...+P(A_n)+...$

Theorem 1: The probability of the impossible event is zero, i.e., if ϕ is the subset (event) containing no sample point, P(ϕ)=0.

Proof: The certain event S and the impossible event ϕ are mutually exclusive.

Hence $P(S \cup \phi)=P(S)+P(\phi)$ [axiom (iii)]

But $S \cup \phi = S$.

Therefore, $P(S)=P(S)+P(\phi)$

Hence $P(\phi)=0.$

Theorem 2: If $\overline{A}$ is the complementary event of A, $P(\overline{A})=1-P(A) \le 1.$

Proof: A and $\overline{A}$ are mutually exclusive events, such that $A \cup \overline{A} = S$

Therefore, $P(A \cup \overline{A})=P(S)$ =1 (Since axiom (ii))

i.e., $P(A)+P(\overline{A})=1.$

Therefore, $P(\overline{A})=1-P(A)$

Since $P(A) \ge 0,$ it follows that $P(\overline{A}) \le 1.$

Theorem 3: If $B \subset A then\ P(B) \le P(A).$

Proof: B and $A\overline{B}$ are mutually exclusive events such that $B \cup A\overline{B} = A$.

Therefore, $P(B \cup A\overline{B})=P(A)$

i.e., $P(B)+P(A\overline{B})=P(A)$ [axiom (iii)]

Therefore, $P(B) \le P(A).$

Theorem 4: Addition theorem of probability

Statement: For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B).$

Proof: Since $(A \cup B) = A \cup (A' \cap B)$ here A and $(A' \cap B)$ are mutually exclusive.

$P(A \cup B) = P[A \cup (A' \cap B)] \ldots (1)$

$= P(A) + P(A' \cap B)$

Again $B = (A \cap B) \cup (A' \cap B)$

Here $(A \cap B) \& (A' \cap B)$ are mutually exclusive events.

$P(B) = P[(A \cap B) \cup (A' \cap B)] \ldots (2)$

$= P(A \cap B) + P(A' \cap B)$

Therefore $P(A' \cap B) = P(B) - P(A \cap B)$

From (1), $P(A \cup B) = P(A) + P(B) - P(A \cap B).$

Conditional Probability: The Conditional probability of an event B, assuming that the event A has happened, is denoted by P(B/A) and defined as, $P(B/A) = \frac{P(A \cap B)}{P(A)}$, provided $P(A) \neq 0.$

Rewriting the definition of conditional probability, we get $P(A \cap B) = P(A) \times P(B/A).$ [Product theorem of probability]

Properties:

1. If $A \subset B, P(B/A) = 1, Since\ A \cap B = A.$
2. If $B \subset A, P(B/A) \geq P(B), Since\ A \cap B = B,$ and $\frac{P(B)}{P(A)} \geq P(B), as\ P(A) \leq P(S) = 1.$
3. If A and B are mutually exclusive events, P(B/A)=0, since $P(A \cap B) = 0$
4. If P(A)>P(B), P(A/B)>P(B/A).
5. If $A_1 \subset A_2, P(A_1/B) \leq P(A_2/B)$.

Independent Events: A set of events is said to be independent if the occurrence of any one of them does not depend on the occurrence or non-occurrence of the others.

The product theorem can be extended to any number of independent events: $A_1, A_{2,\ldots,} A_n$ are n independent events. $P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \times P(A_2) \times \ldots \times P(A_n)$, when this condition is satisfied, the events $A_1, A_{2,\ldots,} A_n$ are also said to be totally independent. A set of events $A_1, A_{2,\ldots,} A_n$ is said to be mutually independent if the events are totally independent when considered in sets of 2,3,. . . n events.

Theorem 5: If the events A and B are independent, then so are $\overline{A} \& \overline{B}.$

Proof. $P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$

$= 1 - [P(A) + P(B) - P(A \cap B)]$ (By addition theorem)

$= 1 - P(A) - P(B) + P(A) \times P(B)$ {since A and B are independent)

$$= [1 - P(A)] - P(B)[1 - P(A)]$$

$$= P(\overline{A}) \times P(\overline{B})$$

Example 1: In how many different ways can the director of a research laboratory choose two chemists from among seven applicants and three physicists from among nine applicants?

Solution:

The two chemists can be chosen in ${}^{7}C_{2}$ =21 ways

The three physicists can be chosen in ${}^{9}C_{3}$ = 84 ways

Then these two things can be done in 21 x 84 = 1764 ways.

Example 2: What is the probability that a non-leap year contains 53 Sundays?

Solution:

A non-leap year consists of 365 days, of these there are 52 complete weeks and 1 extra day. That day may be any one of the 7 days. So already we have 52 Sundays. For one more Sunday, the probability that getting a one more Sunday is 1/7.

Hence the probability that a non-leap year contains 53 Sundays is 1/7.

Example 3: A bag contains 7 white, 6 red and 5 black balls. Two balls are drawn at random. Find the probability that they will both the white?

Solution:

Given that Balls White(7), Red(6) & Black(5), total 18 balls.

Two balls are drawn at random from 18 balls in ${}^{18}C_{2}$ ways

Two white balls are drawn at random from 7 balls in ${}^{7}C_{2}$ ways.

Hence the required probability = $({}^{7}C_{2})/({}^{18}C_{2}) = 21/153.$

Example 4 : Determine the probability that for a non-defective bolt will be found if out of 600 bolts already examined 12 were defective.

Solution:

Given that out of 600 bolts 12 were defective.

Therefore, probability that a defective bolt will be found = $\frac{12}{600} = \frac{1}{50}$

Therefore, Probability of getting a non-defective bolt = $1 - \frac{1}{50} = \frac{49}{50}$.

Example 5: A fair coin is tossed 4 times. Define the sample space corresponding to this experiment. Also give the subsets corresponding to the following events and find the respective probabilities:

a).More heads than tails are obtained.

b).Tails occur on the even numbered tosses.

Solution:

S= {HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT}

a). Let A be the event is which more heads occur than tails

Then A= {HHHH, HHHT, HHTH. HTHH,THHH}

b).Let B be the event is which tails occur is the second and fourth tosses.

Then B= {HTHT, HTTT,TTHT,TTTT}

$$P(A)=\frac{n(A)}{n(S)}=\frac{5}{16};\ P(B)=\frac{n(B)}{n(S)}=\frac{1}{4}.$$

Example 6: A box contains 4 bad & 6 good tubes. Two are drawn out from the box at a time. One of them is tested and found to be good. What is probability that the other one is also good?

Solution:

Let A = one of the tubes drawn is good and B = the other tube is good .

$P(A\cap B)$ = P(both tubes drawn are good)

$$=\frac{{}^{6}C_{2}}{{}^{10}C_{2}}=\frac{1}{3}$$

Knowing that one tube is good, the conditional probability that the other tube is also good is required, i.e., P(B/A) is required.

By definition, $$P(B/A)=\frac{P(A\cap B)}{P(A)}=\frac{1/3}{6/10}=\frac{5}{9}.$$

Example 7: In a shooting test, the probability of hitting the target is ½ for A , 2/3 for B , 3/4 for C. If all of them five at the target, find the probability that

i). none of them hits the target.

ii). Atleast one of them hits the target.

Solution:

Let A = event of A hitting the target.

$$P(\overline{A})=\frac{1}{2},P(\overline{B})=\frac{1}{3},P(\overline{C})=\frac{1}{4}.$$

$$P(\overline{A}\cap\overline{B}\cap\overline{C})=P(\overline{A})\times P(\overline{B})\times P(\overline{C})$$ (by independence)

i.e., P(none hits the target) = $$\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}=\frac{1}{24}$$

P(atleast one hits the target) = 1 – P(none hits the target)

$$=1-\frac{1}{24}=\frac{23}{24}.$$

Example:8

Three coins are tossed together find they are exactly 2 head?

Solution:

Total no. of chances by throwing 3 coins are n(S)= 8.

The event A to get exactly 2 heads are A = {HHT, THH,HTH}

n(A)= 3

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

Example:9

A bag contains 4 red, 5 white and 6 black balls. What is the probability that 2 balls drawn are red and black?

Solution:

Given that Balls White(5), Red(4) & Black(6), total 15 balls.

Two balls are drawn at random from 15 balls in $^{15}C_2$ ways

n(A)= $4C_1X\,6C_1$, Hence the required probability = $\frac{4C_1 X 6C_1}{15C_2} = \frac{8}{35}$

Example :10

A bag contains 3 red and 4 white balls. Two draws are made without replacement.

What is the probability that both balls are red

Solution:

Total no. of balls = 3Red + 4 White = 7 balls

P(Drawing a red ball in the first drawn is red) = $P(A) = \frac{3}{7}$

P(Drawing a red ball in the second drawn is red) = $P(B/A) = \frac{2}{6}$

$$P(A \quad B) = P(A)P(B)$$

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

$$P(A \quad B) = P(A)P(B/A)$$

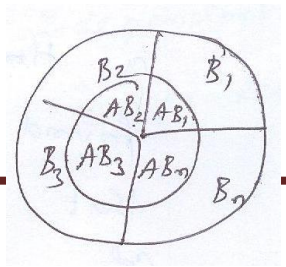
$$= \frac{1}{7}$$

Theorem of Total Probability

Statement: If $B_1, B_{2,\ldots,} B_n$ be a set of exhaustive and mutually exclusive events, and A is another event associated with (or caused by) B_i, then $P(A) = \sum_{i=1}^{n} P(B_i)\,P(A/B_i)$

Proof. The inner circle represents the event A. A can occur along with (or due to) $B_1, B_{2,\ldots,} B_n$ that are exhaustive and mutually exclusive.

Therefore, $AB_1, AB_{2,\ldots,} AB_n$ are also mutually exclusive.



Therefore, $A = AB_1 + AB_2 + \ldots + AB_n$ (by addition theorem)

Hence $P(A) = P(\sum AB_i)$

$= \sum P(AB_i)$ (since $AB_1, AB_2, \ldots, AB_n$ are mutually exclusive)

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

Baye's theorem on Probability (or) Rule of inverse probability

Statement: If $B_1, B_2, \ldots, B_n$ be a set of exhaustive and mutually exclusive events associated with a random experiment and A is another event associated with (or caused by) B_i, then

$$P(B_i/A) = \frac{P(B_i) \times P(A/B_i)}{\sum_{i=1}^{n} P(B_i) \times P(A/B_i)}, \; i = 1,2,\ldots,n$$

Proof. Since by product theorem, $P(A \cap B_i) = P(B_i) \times P(A/B_i)$ . . . (1)

or $P(A \cap B_i) = P(A)\ P(B_i/A)$ . . . (2)

From (1) and (2), $P(A)P(B_i/A) = P(B_i)\ P(A/B_i)$

$$P(B_i/A) = \frac{P(B_i)\ P(A/B_i)}{P(A)} \quad \ldots (3)$$

Therefore from total probability, $P(A) = \sum_{i=1}^{n} P(B_i) P(A/B_i)$ substitute in (3), we get

$$P(B_i/A) = \frac{P(B_i) \times P(A/B_i)}{\sum_{i=1}^{n} P(B_i) \times P(A/B_i)}, \; i = 1,2,\ldots,n$$

Example 11: A bag contains 5 balls and it is not known how many of them are white. Two balls are drawn at random from the bag & they are note to be white. What is the chance the all the balls in the bag are white?

Solution:

Since 2 white balls have been drawn out, the bag must have contained 2, 3, 4, or 5 white balls.

Let B_1 = Event of the bag containing 2 white balls.

B_2 = Event of the bag containing 3 white balls.

B_3 = Event of the bag containing 4 white balls.

B_4 = Event of the bag containing 5 white balls.

Let A = Event of drawing 2 white balls.

$$P(A/B_1) = \frac{^2C_2}{^5C_2} = \frac{1}{10}, \quad P(A/B_2) = \frac{^3C_2}{^5C_2} = \frac{3}{10}$$

$$P(A/B_3)=\frac{^4C_2}{^5C_2}=\frac{3}{5},\ P(A/B_4)=\frac{^5C_2}{^5C_2}=1$$

Since the number of white balls in the bag is not known, B_i's are equally likely.

Therefore $P(B_1)=P(B_2)=P(B_3)=P(B_4)=\frac{1}{4}$

By Baye's theorem,

$$P(B_4/A)=\frac{P(B_4)\times P(A/B_4)}{\sum_{i=1}^{4}P(B_i)\times P(A/B_i)}=\frac{\frac{1}{4}\times 1}{\frac{1}{4}\times\left(\frac{1}{10}+\frac{3}{10}+\frac{3}{5}+1\right)}=\frac{1}{2}.$$

Example 12: There are 3 true coins and 1 false coin with 'head' on both sides. A coin is closer at random and tosses 4 times, If 'head' occurs are the 4 times, What is the probability that the false coin has been chosen and used?

Solution:

P(T) = P(the coin is a true coin) = 3/4

P(F) = P(the coin is a false coin) = 1/4

Let A = Event of getting all heads is 4 tosses,

Then, P(A/T) = ½ * ½ * ½ * ½ = 1/16 and P(A/F) = 1

$$P(F/A)=\frac{P(F)\times P(A/F)}{P(F)\times P(A/F)+P(T)\times P(A/T)}=\frac{\frac{1}{4}\times 1}{\frac{1}{4}\times 1+\frac{3}{4}\times\frac{1}{16}}=\frac{16}{19}.$$

By Baye's theorem,

Example 13:

There are three bags , bag one contains 3 white balls , 2 red balls and 4 black balls. Bag two contains 2 white balls, 3 red balls and 5 black balls. Bag three contains 3 white balls, 4 red balls and 2 black balls. One bag is chosen at random and from it 3 balls were drawn out of which 2 balls were white and 1 is red. What is the probability that it is drawn from bag one, two and three?

Solution:

Selection of bags are mutually exclusive events. The selection of the 2 white and 1 red ball is an independent event.

$P(B_1)=P(B_2)=P(B_3)=1/3$

$P(A/B_1)$ = P(Bag 1 selected from 2W&1R ball chosen)

$$=\frac{3C_2 X 2C_1}{9C_3}$$

=0.07

$P(A/B_2)$ = P(Bag 2 selected from 2W&1R ball chosen)

$$=\frac{2C_2 X 3C_1}{10C_3}$$

= 0.025

$P(A/B_3)$ = P(Bag 3 selected from 2W&1R ball chosen)

$$= \frac{3C_2 X 4C_1}{9C_3}$$
$$= 0.14$$

By using Baye's theorem we have

$P(B_i)$	$P(A/B_i)$	$P(B_i)\ P(A/B_i)$
1/3	0.07	0.0233
1/3	0.025	0.0083
1/3	0.14	0.0466
	$\sum P(B_i)\ P(A/B_i)$	0.0782

$P(B_1/A)$ = P(The balls selected from the first bag)
$$= \frac{0.0233}{0.0782}$$
$$= 0.29$$

$P(B_2/A)$ = P(The balls selected from the second bag)
$$= \frac{0.008}{0.0782}$$
$$= 0.102$$

$P(B_3/A)$ = P(The balls selected from the third bag)
$$= \frac{0.046}{0.0782}$$
$$= 0.58$$

Exercise:

1. In a bolt factory machines A,B,C manufactures 25%,35% and 40% of the total respectively. Out of their output 5%,4% and 2% are defective bolts respectively. A bolt is drawn at random and is found to be defective. What are the probabilities that it was manufactured by the machines A,B and C respectively?

2. A bag contains five balls and it is not known how many of them are white. Two balls are drawn at random from the bag and they are found to be white. What is the probability that all the balls in the bag are white?

RANDOM VARIABLES

Definition: A real-valued function defined on the outcome of a probability experiment is called a random variable. A Random variable (RV) is a rule that assigns a numerical value to each possible outcome of an experiment.

1. Discrete Random Variables.
2. Continuous Random Variables

Probability distribution function of X: If X is a random variable, then the function F(x) defined by $F(x) = P\{X \le x\}$ is called the distribution function of X.

1. **Discrete Random Variable:** A random variable whose set of possible values is either finite or countable infinite is called discrete random variable.

Probability Mass Function (pmf): If X is a discrete variable, then the function $p(x) = P[X = x]$ is called the pmf of X. It satisfies two conditions

i) $p(x_i) \geq 0$

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

Cumulative distribution [discrete R.V] or distribution function of X: The cumulative distribution F(x) of discrete random variable X with probability f(x) is given by

$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)\ for -\infty < x < \infty$

Properties of distribution function:

1. $F(-\infty) = 0$
2. $F(\infty) = 1$
3. $0 \leq F(x) \leq 1$
4. $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$
5. $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) + P[X = x_1]$
6. $P(x_1 < X < x_2) = F(x_2) - F(x_1) - P[X = x_2]$
7. $P(x_1 \leq X < x_2) = F(x_2) - F(x_1) - P[X = x_2] + P[X = x_1]$

Results:

1. $P(X \leq \infty) = 1$
2. $P(X \leq -\infty) = 0$
3. $P(X > x) = 1 - P[X \leq x]$
4. $P(X \leq x) = 1 - P[X > x]$

Example: 14

A continuous random variable 'X' has a probability density function f(x) = K, $0 \leq x \leq 1$. Find 'K'.

Solution:

Given $f(x) = k, 0 \leq x \leq 1$

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

$$\int_{0}^{\infty} kdx = 1$$

k=1

Example 15: A R.V X has the following probability distribution.

x:	-2	-1	0	1	2	3
p(x):	0.1	k	0.2	2k	0.3	3k

Find (1) The value of k, (2) Evaluate P(X<2) and P(-2<X<2) .

Solution:

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

0.1+k+0.2+2k+0.3+3k = 1

K = 1/15.

(2) P[X<2] = P[x=-2] + P[x=-1] + P[x=0] + P[x=1]

= 0.1 + 1/15 + 0.2 + 2/15

= ½

P[-2 < X < 2] = P[x=-1]+ P[x=0]+ P[x=1]

= 1/15 + 0.2 + 2/15 = 2/5

Example 16:

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 ≥ 3) and P (0 < X< 5).

iii) Find the distribution function of X.

Solution:

i) To find 'a' value:

Given discrete random variable, $\sum_{i=1}^{\infty} p(x_i) = 1$

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

81a=1

a=1/81

ii) To find P(X<3):

P(X<3) = P(X=0)+P(X=1)+P(X=2)

=a+3a+5a

=9a

=1/9

iii) To find $P(X \geq 3)$:

$P(X \geq 3) = 1 - P(X < 3)$

=1-1/9 =8/9

iv) To find $P(0 < X < 5)$:

$P(0 < X < 5) = P(X = 1) + ...P(X = 4)$

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

= 24/81

v) To find the distribution function of X:

Value of x	0	1	2	3	4	5	6	7	8
P(x)	a	3a	5a	7a	9a	11a	13a	15a	17a
P(x)	1/81	3/81	5/81	7/81	9/81	11/81	13/81	15/81	17/81
F(x)	1/81	4/81	9/81	16/81	25/81	36/81	49/81	64/81	1

Example 17: A R.V X has the following function:

X: 0 1 2 3 4 5 6 7

P(X): 0 k 2k 2k 3k k^2 $2k^2$ $7k^2+k$

(a) find k (b) Evaluate P[X<6], P[x≥6], (c) Evaluate P[1.5<X<4.5 / X>2] (d) Find P[X<2], P[X>3], P[1<X<5].

Solution:

(a). Since $\sum_{i=1}^{n} p(x_i) = 1$

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

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

K = -1 or 1/10 (since k=-1 is not permissible, P(X)≥0)

Hence k = 1/10.

(b). P[x≥6] = P[X=6] + P[X=7]

$= 2k^2+7k^2+k$

= 2/100 + 7/100 + 1/10 = 19/100

P[X<6] = 1 - P[x≥6]

= 1 – 19/100

= 81/100

(c). P[1.5<X<4.5 / X>2] $= \frac{p[(1.5 < x < 4.5) \cap x > 2]}{p(x > 2)}$ (by conditional probability)

$= \frac{p[2 < x < 4.5]}{1 - p(x \le 2)}$

$= \frac{p(3) + p(4)}{1 - [p(0) + p(1) + p(2)]}$

$= \frac{\frac{2}{10} + \frac{3}{10}}{1 - \left[0 + \frac{1}{10} + \frac{2}{10}\right]} = \frac{\frac{5}{10}}{\frac{7}{10}} = \frac{5}{7}$

(d). p(X<2) = p[x=0] + p[x=1]

= 0 + k = k = 1/10

P(X>3) = 1 – p(x≤3)

= 1 – [p(x=0)+p(x=1)+p(x=2)+p(x=3)]

= 1 – [0+k+2k+2k]

= ½

P(1<x<5) = p(x=2) + p(x=3)+ p(x=4)

= 2k + 2k + 3k

= 7/10

Example 18: If the R.V. X takes the values 1,2,3 and 4 such that 2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4). Find the probability distribution and cumulative distribution function of X.

Solution:

Since X is a discrete random variable.

Let 2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4) = k

2P(X = 1) = k implies that P(X = 1) = k/2

3P(X = 2) = k implies that P(X = 2) = k/3

P(X = 3) = k

5P(X = 4) = k implies that P(X = 4) = k/5

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

i.e., k/2 + k/3 + k + k/3 = 1

k[1/2 + 1/3 + 1 + 1/5] = 1

Therefore k = 30/61

x_i	$p(x_i)$	F(X)
1	P(1) = k/2 = 15/61	F(1) = p(1) = 15/61
2	P(2) = k/3 = 10/61	F(2) = F(1) + p(2) = 15/61 + 10/61 = 25/61
3	P(3) = k = 30/61	F(3) = F(2)+p(3) = 25/61 + 30/61 = 55/61
4	P(4) = k/5 = 6/61	F(4) = F(3)+p(4) = 55/61 + 6/61 = 61/61 = 1

Example 19: A discrete random variable X has the following probability mass function:

X	0	1	2	3	4	5	6	7
P(X)	0	a	2a	2a	3a	a^2	$2a^2$	$7a^2+a$

Find (i) the value of 'a' (ii) $P(X<6),\ P(X\geq 6)$ (iii) $P(0<X<5)$ (iv) the distribution function of X (v) If $P(X\leq x)>1/2$, find the minimum value of X.

Solution:

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

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

$10a^2 + 9a - 1 = 0$

$a = -1$ or $1/10$ (since $a=-1$ is not permissible, $P(X)\geq 0$)

Hence $a = 1/10$.

(ii). $P[x\geq 6] = P[X=6] + P[X=7]$

$= 2a^2+7a^2+a$

$= 2/100 + 7/100 + 1/10 = 19/100$

(iii). $P[X<6] = 1 - P[x\geq 6]$

$= 1 - 19/100$

$= 81/100$

(iv). To find P(0<X<5):

$P(0<X<5) = P(X=1)+.....P(X=4)$

$= a+2a+2a+3a$

$= 8a = 8/10$

(v). To find distribution function of X :

x	0	1	2	3	4	5	6	7
P(x)	0	a	2a	2a	3a	a^2	$2a^2$	$7a^2+a$
F(x)	0	1/10	3/10	5/10	8/10	81/100	83/100	1

Minimum value of X:

$P(X \leq x)$ $1/2$

The minimum value of X for which $P(X \leq x)$ 0.5, is the x value is 4.

2. **Continuous Random Variables:** A random variable X is said to be continuous if it takes all possible values between certain limits say from real number ‘a’ to real number ‘b’.

Example: The length time during which a vacuum tube installed in a circuit functions is a continuous random variable, number of scratches on a surface, proportion of defective parts among 1000 testes, number of transmitted in error.

Probability density function (pdf): For a continuous R.V X, a probability density function is a function such that (1) $f(x) \geq 0$ (2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (3)

$P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx =$ area under f(x) from a to b for any a and b.

Cumulative distribution function: The Cumulative distribution function of a continuous R.V. X is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt \quad for -\infty < x < \infty.$$

Mean and variance of the Continuous R.V. X: Suppose X is continuous variable with pdf f(x). The mean or expected value of X, denoted as μ or E(X)

$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$. And the variance of X, denoted as V(X) or σ^2 is $E[X^2] - [E(X)]^2$

Example 20: Given that the pdf of a R.V X is f(x)=kx, 0<x<1. Find k and P(X>0.5)

Solution:

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

$$\int_{0}^{1} kx\,dx = 1$$

$$k\left[\frac{x^2}{2}\right]_0^1 = 1$$

K = 2

P(X > 0.5) = $\int_{0.5}^{\infty} f(x)\,dx$

$$= \int_{1/2}^{1} 2x\,dx$$

$$= 2\left[\frac{x^2}{2}\right]_{1/2}^{1}$$

$= 3/4$

Example 21: If $f(x) = \begin{cases} kxe^{-x}, & x > 0 \\ 0, & elsewhere \end{cases}$ is the pdf of a R.V. X. Find k.

Solution:

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

Here $\int_0^\infty kxe^{-x}\,dx = 1$ [since x>0]

$$k\left[x\left(\frac{e^{-x}}{-1}\right) - 1\left(\frac{e^{-x}}{-1}\right)\right]_0^\infty = 1$$

K =1

Example 22: A continuous R.V. X has he density function $f(x) = \frac{k}{1+x^2}, -\infty < x < \infty.$ find the value of k and the distribution function.

Solution:

Given is a pdf $\int_{-\infty}^{\infty} f(x)\,dx = 1$, $f(x) = \frac{k}{1+x^2}, -\infty < x < \infty.$

$$k\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = 1$$

$$2k\int_0^{\infty} \frac{1}{1+x^2}\,dx = 1$$

$$2k\left[\tan^{-1} x\right]_0^{\infty} = 1$$

$$2k\left[\frac{\pi}{2} - 0\right] = 1$$

$$\pi k = 1; k = \frac{1}{\pi}$$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} \frac{1}{\pi}\left(\frac{1}{1+x^2}\right)dx$$

$$= \frac{1}{\pi}\left[\tan^{-1} x\right]_{-\infty}^{x} = \frac{1}{\pi}\left[\tan^{-1} x - \left(\frac{-\pi}{2}\right)\right]$$

$$= \frac{1}{\pi}\left[\tan^{-1} x + \left(\frac{\pi}{2}\right)\right] \text{ for } -\infty < x < \infty$$

Example:23

A continuous random variable X has a pdf $f(x) = 3x^2,\ 0 \le x \le 1.$ Find a and b such that (i) $P(X \le a) = P(X > a)$ and (ii) $P(X > b) = 0.05.$

Solution:

A continuous random variable X has a pdf $f(x) = 3x^2,\ 0 \le x \le 1.$

i) To find $P(X \le a) = P(X \quad a)$

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

$$\int_0^1 3x^2 dx = 1$$

$P(X \le a) = P(X \quad a)$

Since $P(X \le a) = \frac{1}{2} = 0.5$,

$\int_0^a f(x)dx = \frac{1}{2}$, $\int_0^a 3x^2 dx = a^3 = \frac{1}{2}$

a = 0.7937

ii) To find $P(X \quad b) = 0.05$

$\int_b^1 f(x)dx = 0.05$, $\int_b^1 3x^2 dx = 1 - b^3 = 0.05$

$b^3 = 0.95$

$b = (0.95)^{1/3}$

$$f(x) = \begin{cases} ax, & 0 \le x \le 1 \\ a, & 1 \le x \le 2 \\ 3a - ax, & 2 \le x \le 3 \\ 0, & otherwise \end{cases}$$

Example 24: If the density function of a continuous R.V. X is given by

(1) Find the value of a.

(2) The cumulative distribution function of X.

(3) If x_1, x_2, x_3 are 3 independent observations of X. What is the probability that exactly one of these 3 is greater than1.5?

Solution:

(1) Since f(x) is a pdf, then $\int_{-\infty}^{\infty} f(x)\,dx = 1$

i.e., $\int_0^3 f(x)\,dx = 1$

i.e., $\int_0^1 ax\,dx + \int_1^2 a\,dx + \int_2^3 (3a - ax)\,dx = 1$

a = ½

(2). (i) If x<0 then F(x) = 0

(ii) If $0 \le x \le 1 then\ F(x) = \int_0^x ax\,dx = \int_0^x \frac{x}{2}dx$

$= \frac{x^2}{4}$

$(iii)\ If\ 1 \le x \le 2 then\ F(x) = \int_{-\infty}^{x} f(x)\,dx$

$= \int_0^1 ax\,dx + \int_1^x a\,dx$

$= \frac{x}{2} - \frac{1}{4}$

$$(iv)\ If\ 2 \le x \le 3\ then\ F(x) = \int_{-\infty}^{x} f(x)\,dx$$
$$= \int_{0}^{1} ax\,dx + \int_{1}^{2} a\,dx + \int_{2}^{x} (3a - ax)\,dx$$
$$= \frac{3x}{2} - \frac{x^2}{4} - \frac{5}{4}$$

(v) If x>3, then $F(x) = \int_{-\infty}^{x} f(x)\,dx$

$$= \int_{0}^{1} ax\,dx + \int_{1}^{2} a\,dx + \int_{2}^{3} (3a - ax)\,dx + \int_{3}^{x} f(x)\,dx$$
$$= 1$$

(3). $$P(X > 1.5) = \int_{1.5}^{3} f(x)\,dx = \int_{1.5}^{2} \frac{1}{2}\,dx + \int_{2}^{3} \left(\frac{3}{2} - \frac{x}{2}\right) dx$$
$$= ½$$

Choosing an X and observing its value can be considered as a trail and X>1.5 can be considered as a success.

Therefore, p=1/2, q=1/2.

As we choose 3 independent observation of X, n = 3.

By Bernoulli's theorem, P(exactly one value > 1.5) = P(1 success)

$$= {}^{3}C_{1} \times (p)^{1} \times (q)^{2} = \frac{3}{8}.$$

Example 25: A RV X has the following distribution

X	-2	-1	0	1	2	3
P(X)	0.1	k	0.2	2k	0.3	3k

(a) find k (b) Evaluate P(X<2) & P(-2<X<2)

Solution:

(a) Σ P(X)=1

6K+0.6=1

K=1/15

Since the distribution is

X	-2	-1	0	1	2	3
P(X)	1/10	1/15	1/5	2/15	3/10	1/5

(b) P(X<2) = P(X=-2) + P(X=-1) + P(X=0) + P(X=1)

= 1/10 + 1/15 + 1/5 + 2/15 =1/2

& P(-2<X<2) = P(X=-1) + P(X=0) + P(X=1)

$= 1/15 + 1/5 + 2/15 = 2/5$

Example:26

A continuous random variable X is having the probability density function

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

Find the cumulative distribution function of x.

Solution:

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

To find cumulative distribution function of x:

i) If $0<x<1$ $\quad F(x) = \int_{-\infty}^{x} f(x)dx$

$$= \int_{0}^{x} x dx = \frac{x^2}{2}$$

ii) If $1<x<2$, $\quad F(x) = \int_{-\infty}^{x} f(x)dx$

$$= \int_{0}^{1} x dx + \int_{1}^{x} (2-x)dx$$

$$= 2x - \frac{x^2}{2} - 1$$

iii) If $x>2$, $F(x) = \int_{-\infty}^{x} f(x)dx$

$$= \int_{0}^{1} x dx + \int_{1}^{2} (2-x)dx$$

$$= 1$$

The cumulative distribution function of x is

$$F(x)=\begin{cases} \frac{x^2}{2}, & 0<x<1 \\ 2x-\frac{x^2}{2}-1, & 1<x<2 \\ 1, & x>2 \end{cases}$$

	Questions	**OPT 1**	**OPT 2**	**OPT3**
No.	**Questions**	**OPT 1**	**OPT 2**	**OPT3**
1	A numerical measure of uncertainty is practiced by the important branch of statistics called _____	Theory of mathematics	Theory of physics	Theory of statistics
2	If P(A) is 1, the event A is called a _____	Cases	Trial	Certain Event
3	p + q = _____, here p is success and q is failure events	7	9	1
4	In rolling of single die, the chance of getting 2,4,6 (even numbers) are _____	simple	Compound event	Certain event
5	The set of all possible outomes of an activity is the _____	sample space	event	independent
6	Events that cannot happen together are called_____	mutually exculsive	event	exclusive
7	If one event is unaffected by the outcome of another event, the two events are said to be _____	dependent	independent	mutually exclusive
8	If P(A or B)=P(A), then_____	A and B are mutually exclusive	venn diagram	P(A)+P(B)
9	If P(X≤ x)=_____	1-P(X>x)	1	0
10	If the outcome of one event does not influence another event, then the two events are_____	mutually exculsive	dependent	independent
11	If P(A)=0.9, P(B/A)=0.8, find P (A∩B) = _____	0.72	0.17	0.1
12	If P(X>x)=_____	1-P(X≤x)	P(X≤x)	1
13	For a discrete random variable, the probability density function represents the _____	probability mass function	probability distribution function	probability density function
14	Why are the events of a coin toss mutually exclusive_____	the out come of any toss is not affected by the out come of those preceding	both a head and a tail cannot turn up on any one toss	the probabiity of getting a head and the probability of getting a tail

OPT 4	**OPT 5**	**OPT 6**	**ANSWERS**
OPT 4	**OPT 5**	**OPT 6**	**ANSWERS**
Theory of probability			Theory of probability
3			Certain Event
impossible event			1
Theory of probability			Compound event
mode			sample space
11			mutually exculsive
deviation			independent
conditional probability			A and B are mutually exclusive
P(X>x)			1-P(X>x)
random variable			independent
0.86			0.72
0			1-P(X≤x)
zero			probability mass function
1			both a head and tail cannot turn up on any one toss

15	What is the probabiity that a ball drawn at random from the urn is blue_____	0.1	0.4	0.6
16	What is the probability of getting an even number when a die is tossed___	1/3	1/2	1/6
17	What is the probability of getting more than 2 when a die is tossed___	1/3	1/2	2/3
18	The probability of drawing a spade from a pack of cards is_____	1/52	1/13	4/13
19	If A and B are independent event P (A)=0.4 and P(B)=0.5 then P (AUB)=	0.7	0.1	0.3
20	What is the probability of getting a sum 9 from two throws of a dice?	1/6	1/9	8/9
21	Three unbiased coins are tossed. What is the probability of getting at most two heads?	3/4	1/4	3/8
22	A bag contains 6 black and 8 white balls. One ball is drawn at random. What is the probability that the ball drawn is white?	3/4	4/7	1/8
23	Total probability is _____	0	1	-1
24	Probability of a single real value in a continuous random variable is _____	two	three	four
25	A random variable X is_____ if it assumes only discrete values.	spectrum	complex	continuous
26	If P(A)=0.35, P(B)=0.73, P(A∩B) =0.14 find P(A'UB')= _____	0.86	0.115	1.08
27	The classical school of thought on probability assumes that all possible outcomes of an experiment are____	Equally likely	Mutually exclusive	Mutually exclusive and equally likely
28	Rolling of die is a	Trial	Cases	Event
29	If P(A) = 0, the event A is called a	Trial	Impossible ever	Cases
30	The simple probability of an occurrence of an event is called the _____	bayesian probability	joint probability	mariginal probabiity
31	E(ax+b)=	ax+b	aE(x)+b	E(x)
32	The impossible event is	0	1	-1

mode			0.6
1/9			1/2
1/4			2/3
1/4			1/4
0.5			0.7
7/8			1/9
3/7			3/8
none of these			4/7
0.5			1
zero			zero
Random experiment			Random experiment
0.66			0.86
1/9			Mutually exclusive and equally likely
Random experiment			Trial
Event			Impossible event
all of these			marginal probability
None of these			aE(x)+b
0.5			0

33	A density function may correspond to different _____	probability mass function	probability distribution function	probability density function
34	If A and B are independent and P(A) =0.2, P(B)=0.6 find P(A∩B) =	0.12	0.8	0.2

random variable			random variable
0.4			0.12

UNIT-II

RANDOM VARIABLES

Introduction:

In the last chapter, we introduced the concept of a single random variable. We observed that the various statistical averages or moments of the random variable like mean, variance, standard derivation, skewness give an idea about the characteristics of the random variable.

But in many practical problems several random variables interact with each other and frequently we are interested in the joint behavior of the health conditions of a person, doctors measure many parameters like height, weight, blood pressure, sugar level etc. we should now introduce techniques that help us to determine the joint statistical properties of several random variables.

The concepts like distribution function, density function and moments that we defined for single random variable can be extended to multiple random variables also.

Definition:

Let S be the sample space. Let X=X(S) and Y=Y(S) be two functions each assigning a real no. to each outcome $s \in S$. Then (X,Y) is a two dimensional random variable.

Types of random variables:

1. Discrete random variables

2. Continuous random variables

Two dimensional discrete random variables:

If the possible values of (X,Y) are finite or countably infinite then (X,Y) is called a two dimensional discrete random variables when (X,Y) is a two dimensional discrete random variable the possible values of (X,Y) may be represented as (x_i, y_j) i = 1,2,…..n, j=1,2,….m.

Two dimensional continuous random variables:

If (X,Y) can assume all values in a specified region R in the XY plane (X,Y) is called a two dimensional continuous random variables.

Joint distributions – Marginal and conditional distributions:

(i) Joint Probability Distribution:

The probabilities of two events $A = \{X \le x\}$ and $B = \{Y \le y\}$ have defined as functions of x and y respectively called probability distribution function.

$F_X(x) = P(X \le x)$

$F_Y(y) = P(Y \le y)$

Discrete random variable important terms:

i) Joint probability function (or) Joint probability mass function:

For two discrete random variables x and y write the probability that X will take the value of x_i, Y will take the value of y_j as, $P(x,y) = P(X = x_i, Y = y_j)$

ie) $P(X = x_i, Y = y_j)$ is the probability of intersection of events $X = x_i \ \& \ Y = y_j$.

$P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j)$, The function $P(X = x_i, Y = y_j) = P(x_i, y_j)$ is called a joint probability function for discrete random variables X,Y and it is denoted by P_{ij}.

P_{ij} satisfies the following conditions

(i) $P_{ij} > 0$, for every i,j

(ii) $\sum_j \sum_i P_{ij} = 1$

Continuous random variable (or) Joint Probability Density Function:

Definition:

The joint probability density function if (x,y) be the two dimensional continuous random variable then f(x,y) is called the joint probability density function of (x,y) the following conditions are satisfied.

(i) $f(x,y) \geq 0, \ \forall x, y \in R$

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1.$ Where R is a sample space.

Note: $P(a \leq x \leq b, c \leq y \leq d) = \int_a^b \int_c^d f(x,y)dydx$

Joint cumulative distributive function:

If (x,y) is a two dimensional random variable then $F(X,Y) = P(X \leq x, Y \leq y)$ is called a cumulative distributive function of (x,y) the discrete case $F(X,Y) = \sum_j \sum_i P_{ij} = 1$, $y_i \leq y, x_i \leq x.$

In the continuous case $F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x,y)dxdy$

Properties of Joint Probability Distribution function:

1. $O \le P(x_i, y_j) \le 1$
2. $\sum_i \sum_j P(X_i, Y_j) = 1$
3. $P(X_i) = \sum_j P(X_i, Y_j)$
4. $P(y_i) = \sum_j P(X_i, Y_j)$
5. $P(x_i) \ge P(x_i, y_j)$ for any j
6. $P(y_j) \ge P(x_i, y_j)$ for any i

Properties:

1. The joint probability distribution function F_{xy} (X, Y) of two random variable X and Y have the following properties. They are very similar to those of the distribution function of a single random variable.
2. $0 \le f_{XY}(x, y) \le 1$
3. $f_{XY}(\infty, \infty) = 1$
4. $f_{XY}(x, y)$ is non decreasing
5. $f_{XY}(-\infty, y) = F_{xy}(x_1, \infty) = 0$
6. For $x_1 < x_2$ and $y_1 < y_2$, $P(x_1 < X \le x_2, Y \le y_1) = F(x_2, y_1) - F(x_1, y_1)$
7. $P(X \le x_1, y_1 < Y \le y_2) = F(x_1, y_2) - F(x_1, y_1)$

   $P(x_1 < X \le x_2, y_1 < Y \le y_2) = F(x_2, y_2) - F(x_1, y_2) - f(x_2, y_1) + f(x_1, y_1)$
8.
9. $F_Y(y) = F_{XY}(\infty, y) = P(X \le \infty, y \le y) = P(y \le y)$
10. $F_X(x) + F_y(y) - 1 \le F_{XY}(x, y) \le \sqrt{F_X(x) F_Y(y)}$ for all x and y.

These properties can also be easily extended to multi dimensional random variables.

Marginal Probability Distribution function:

(i) Discrete case:

- Let (x,y) be a two dimensional discrete random variable, $P_{ij} = P[X = x_i, Y = y_j]$ then $P(X = x_i) = P_i *$ is called a marginal probability of the function X. Then the collection of the pair $\{x_i, P_i^*\}$ is called a marginal probability of X.
- If $P(Y = y_j) = P_{*j}$ is called a marginal probability of the function Y. Then the collection of the pair $\{y_i, P_{*j}\}$ is called a marginal probability of Y.

(ii) Continuous case:

- The marginal density function of X is defined as $f_x(x) = g(x) = \int_{-\infty}^{\infty} f(x,y)dy$ and
- The marginal density function of Y is defined as $f_y(y) = h(y) = \int_{-\infty}^{\infty} f(x,y)dx$

Conditional distributions:

(i) Discrete case:

- The conditional probability function of X given Y=y $_j$ is given by

$$P[X = x_i / Y = Y_j] = P[X = x_i, Y = y_j] / P[Y = y] = P_{ij} / P * j$$

The set $\{X = x_i, P_{ij} / P * j\}$, I =1,2,3,…is called the conditional probability distribution of X given $Y = y_j$

- The conditional probability function of Y given X=xi is given by

$$P = [Y = y_j / X = x_i] = P[Y = y_j, X = x_i] / P[X = x_i] = P_{ij} / P_i^*$$

The set $\{y_{ij} \quad P / P_i^*\}$, j=1,2,3,…is called the conditional probability distribution of Y given $X = x_i$

(ii) Continuous case:

- The conditional probability density function of X is given by $Y = y_j$ is defined as

$f(x/y) = \frac{f(x,y)}{h(y)}$, where h(y) is a marginal probability density function of Y.

- The conditional probability density function of Y is given by $X = x_i$ is defined as

$f(y/x) = \frac{f(x,y)}{g(x)}$, where g(x) is a marginal probability density function of X.

Independent random variables:

(i) Discrete case:

Two random variable (x,y) are said to be independent if $P(X = x_i \cap Y = y_j) = P(X = x_i)(Y = y_j)$ (ie) $P_{ij} = P_i^* P_{*j}$ for all i ,j.

(ii) Continuous case:

Two random variables (x,y) are said to be independent if $f(x,y) = g(x)h(y)$, where $f(x,y) =$ joint probability density function of x and y,

$g(x) =$ Marginal density function of x,

$h(y) =$ Marginal density function of y.

Marginal Distribution Tables:

Table – I

To calculate marginal distribution when the random variables X takes horizontal values and Y takes vertical values

Y/X	x1	x2	x3	p (y) = p(Y=y)
y1	p11	p21	p31	p(Y=y1)
y2	p12	p22	p32	p(Y=y2)
y3	p13	p23	p33	p(Y=y3)
$P_X(X) = P(x = x)$	$P(x = x1)$	$p(x = x2)$	$p(x = x3)$	

Table – II

To calculate marginal distribution when the random variables X takes vertical values and Y takes horizontal values

Y\X	y1	y2	y3	P_X (x) = P(X=x)
x1	p11	p21	p31	p(X=x1)
x2	p12	p22	p32	p(X=x2)
x3	p13	p23	p33	p(X=x3)
$p(y) = p(y = y)$	$P(y = y1)$	$P(y = y2)$	$P(y = y3)$	

Solved Problems on Marginal Distribution:

Example :1

From the following joint distribution of X and Y find the marginal distribution

X/Y	0	1	2
0	3/28	9/28	3/28
1	3/14	3/14	0
2	1/28	0	0

Solution:

The marginal distribution are given in the table below

Y\X	0	1	2	$P_Y(y)$ = P(Y=y)
0	3/28	9/28	3/28	15/28
1	3/14	3/ 14	0	6/14
2	1/28	0	0	1/28
$P_X(x) = P(Y = y)$	$P_X(0)$ = 5/14	$P_X(1)$ = 15/28	$P_X(2)$ = 3/28	1

The marginal Distribution of X

$P_X(0) = P(X = 0) = p(0,0) + p(0,1) + p(0,2)$ $= 3/28 + 3/14 + 1/28 = 5/14$

$P_X(1) = P(X = 1) = p(1,0) + p(1,1) + p(1,2)$ $= 9/28 + 3/14 + 0 = 15/28$

$P_X(2) = P(X = 2) = p(2,0) + p(2,1) + p(2,2)$ $= 3/28 + 0 + 0 = 3/28$

Marginal probability function of X is $P_x(x) = \begin{cases} 5/14, & x = 0 \\ 15/28, & x = 1 \\ 3/28, & x = 2 \end{cases}$

The marginal distributions are

Y/X	1	2	3	$P_Y(y) = p(y = y)$
1	2/21	3/21	4/21	9/21
2	3/21	4/21	5/21	12/21

	5/21	7/21	9/21	1

The marginal distribution of X

$$P_x(1) = p(1,1) + (1,2) \quad = 2/21 + 3/21$$

$$P_x(1) = 5/21$$

$$P_x(2) = p(2,1) + (2,2) \quad = 3/21 + 4/21$$

$$P_x(2) = 7/21$$

$$P_x(3) = p(3,1) + p(3,2) \quad = 4/21 + 5/21$$

$$P_x(3) = 9/21$$

Marginal probability function of X is, $P_x(x) = \begin{cases} 5/21, & x = 1 \\ 7/21, & x = 2 \\ 9/21, & x = 3 \end{cases}$

The marginal distribution of Y

$$P_Y(1) = p(1,1) + p(2,1) + p(3,1) = 2/21 + 3/21 + 4/21$$

$$P_Y(1) = 9/21$$

$$P_Y(2) = p(1,2) + p(2,2) + p(3,2) \quad = 3/21 + 4/21 + 5/21$$

$$P_Y(2) = 12/21$$

Marginal probability function of Y is $P_Y(y) = \begin{cases} 3/21, & y = 1 \\ 4/21, & y = 2 \end{cases}$

Example :2

From the following table for joint distribution of (X, Y) find

i) $P(X \le 1)$ ii) $P(Y \le 3)$ iii) $P(X \le 1, Y \le 3)$ iv) $P(X \le 1 / Y \le 3)$

v) $P(Y \le 3 / X \le 1)$ vi) $P(X + Y \le 4)$.

X/Y	0	2	3	4	5	6
0	0	0	1/32	2/32	2/32	3/32

1	1/16	1/16	1/8	1/8	1/8	1/8
2	1/32	1/32	1/64	1/64	0	2/64

Solution:

The marginal distributions are

X / Y	1	2	3	4	5	6	$P_X(x) = P(X = x)$
0	0	0	1/32	2/32	2/32	3/32	8/32 P(x=0)
1	1/16	1/16	1/8	1/8	1/8	1/8	10/16 P(x=1)
2	1/32	1/32	1/64	1/64	0	2/64	8/64 P(x=2)
$P_Y(y) = P(Y = y)$	3/32	3/32	11/64	13/64	6/32	16/64	1
	P(Y=1)	P(Y=2)	P(Y=3)	P(Y=4)	P(Y=5)	P(Y=6)	

i) $P(X \le 1)$

$P(X \le 1) = P(X = 0) + P(X = 1)$

$= 8/32 + 10/16$

$P(X \le 1) = 28/32$

ii) $P(Y \le 3)$

$P(Y \le 3) = P(Y = 1) + P(Y = 2) + P(Y = 3)$

$= 3/32 + 3/32 + 11/64$

$P(Y \le 3) = 23/64$

iii) $P(X \le 1, Y \le 3)$

$P(X \le 1, Y \le 3) = P(0,1) + P(0,2) + P(0,3) + P(1,1) + P(1,2) + P(1,3)$

$= 0+0+1/32 + 1/16 + 1/16 + 1/8$

$P(X \le 1, Y \le 3) = 9/32$

iv) $P(X \le 1 / Y \le 3)$

By using definition of conditional probability

$$P[x = x_i / y = y_j] = \frac{P[X = x_i, Y = y_j]}{P[Y = y_j]}$$

The marginal distribution of Y

$P_Y(0) = P(Y = 0) = p(0,0) + p(1,0) + p(2,0) = 3/28+9/28+3/28 = 15/28$

$P_y(1) = P(y = 1) = p(0,1) + p(1,1) + p(2,1) = 3/14+3/14+0= 3/7$

$P_y(2) = P(y=2) = p(0,2) + p(1,2) + p(2,2)$ = 1/28+0+0 = 1/2

Marginal probability function of Y is $P_y(Y) = \begin{cases} 15/28, & y=0 \\ 3/7, & y=1 \\ 1/28, & y=2 \end{cases}$

Example 3:

The joint distribution of X and Y is given by f(X, Y) =X+Y/21, x=1,2,3 y=1,2.Find the marginal distributions.

Solution:

Given f (X, Y) =X+Y/21, x=1, 2, 3 y=1,2

f (1,1) = 1+1/21 =2/21 =P(1,1)

f (1,2) = 1+2/21 =3/21 =P(1,2)

f (2,1) = 2+1/21 =3/21 =P(2,1)

f (2,2) = 2+2/21 =4/21 =P(2,2)

f (3,1) = 3+1/21 =4/21 =P(3,1)

f (3,2) = 3+2/21 =5/21 =P(3,2)

$P[X \le 1/Y \le 3] = \frac{P[X \le 1, Y \le 3]}{P[Y \le 3]} = \frac{9/23}{23/64}$

$P[X \le 1/Y \le 3]$ = 18/32

v) $P[Y \le 3/X \le 1]$

$P[Y \le 3/X \le 1] = \frac{P[X \le 3, Y \le 1]}{P[Y \le 1]} = \frac{9/23}{7/8}$

P[Y≤3/ X≤1]= 9/28

vi) $P(X+Y \le 4)$

P(X + Y≤4)) = P(0,1)+P(0,2)+P(0,3)+ P(0,4)+ P(1,1)+
P(1,2)+P(1,3)+P(2,1)+ P(2,2)
= 0+0+1/32 +2/32 + 1/16 + 1/16 + 1/8+1/32 +1/32

P(X + Y≤4) = 13/32

Example : 4

If the joint P.D.F of (X,Y) is given by p(X,Y)=K(2x+3y),x=0,1,2, y=1,2,3,. Find all the marginal probability distribution .Also find the probability of (X+Y) and P(X+Y >3).

Solution:

Given P(X,Y)= K(2x+3y)

P(0,1)= K(0+3) = 3K

P(0,2)= K(0+6) = 6K

P(0,3)= K(0+9) = 9K

P(1,1)= K(2+3) = 5K

P(1,2)= K(2+6) = 8K

P(1,3)= K(2+9) = 11K

P(2,1)= K(4+3) = 7K

P(2,2)= K(4+6) = 10K

P(2,3)= K(4+9) = 13K

To find K:

The marginal distribution is given in the table.

Y\X	0	1	2	$P_Y(y) = P(Y = y)$
1	3K	5K	7K	15K
2	6K	8K	10K	24K
3	9K	11K	13K	33K
PX(x)=P(X=x)	18K	24K	30K	72K

Total Probability =1

72K = 1

K = 1/72

Marginal probability of X & Y:

Substituting K = 1/72 in the above table, we get

Y\X	0	1	2	P_Y (y)=P(Y=y)

1	3/72	5/72	7/72	5/24
2	6/72	8/72	10/72	1/3
3	9/72	11/72	13/72	11/24
P_X (x)=P(X=x)	1/4	11/72	5/12	1

From table, $P_x(0) = 1/4$, $p_x(1) = 1/3$, $p_x(2) = 5/12$

Marginal probability function of x is , $$P_x(X) = \begin{cases} 1/4, x = 0 \\ 1/3, x = 1 \\ 5/2, x = 2 \end{cases}$$

From table, $p_y(1) = 5/24$, $P_y(2) = 1/3$, $P_Y(3) = 11/24$

$$\mathrm{P_Y(Y)} = \begin{cases} 5/24, \mathrm{Y} = 1 \\ 11/24, \mathrm{y} = 2 \end{cases}$$

Marginal Probability function of Y is ,

Example :5

From the following table for joint distribution of (X, Y) find

The marginal distributions are

Y/X	1	2	3	$P_Y(y) = P(Y = y)$
1	2/21	3/21	4/21	9/21
2	3/21	4/21	5/21	12/21
$P_X(x) = P(X = x)$	5/21	7/21	9/21	1

The marginal distribution of X

$P_X(1) = P(1,1)+P(1,2) = 2/21 + 3/21 = P^X(3)=9/21$

$P_X(2) = P(2,1)+P(2,2) = 3/21 + 4/21 = Py(2)=7/21$

$P_X(3) = P(3,1)+P(3,2) = 4/21 + 5/21 = P_X(3)=9/21$

Marginal probability function of X is $P_x(x) = \begin{cases} 5/21, x=1 \\ 7/21, x=2 \\ 9/21, x=3 \end{cases}$

The marginal distribution of Y

$$P_y(1) = P(1,1) + P(2,1) + P(3,1)$$
$$= 2/21 + 3/21 + 4/21 = 9/21$$
$$P_y(2) = P(1,2) + P(2,2) + P(3,2)$$
$$= 3/21 + 4/21 + 5/21 = 12/21$$

Marginal probability function of Y is $P_Y(y) = \begin{cases} 3/21, y=1 \\ 4/21, y=2 \end{cases}$

Exercises:

1. Given is the joint distribution of X and Y

Y/X	0	1	2
0	0.02	0.08	0.10
1	0.05	0.20	0.25
2	0.03	0.12	0.15

Obtain 1) Marginal Distribution.

2) The conditional distribution of X given Y =0.

2. The joint probability mass function of X & Y is

X/Y	0	1	2
0	0.10	0.04	0.02
1	0.08	0.20	0.06
2	0.06	0.14	0.30

Find the M.D.F of X and Y. Also $(X \le 1, Y \le 1)$ and check if X & Y are independent.

3. Let X and Y have the following joint probability distribution

Y/X	2	4
1	0.10	0.15

3	0.20	0.30
5	0.10	0.15

Show that X and Y are independent.

4. The joint probability distribution of X and Y is given by the following table.

X/Y	1	3	9
2	1/8	1/24	1/12
4	¼	¼	0
6	1/8	1/24	1/12.

i) Find the probability distribution of Y.

ii) Find the conditional distribution of Y given X=2.

ii) Are X and Y are independent.

5. Given the following distribution of X and Y. Find

i) Marginal distribution of X and Y.
ii) The conditional distribution of X given Y=2.

X/Y	-1	0	1
0	1/15	2/15	1/15
1	3/15	2/15	1/15
2	2/15	1/15	2/15

Example : 6

If the joint probability density function of (X, Y) is given by $f(x,y)=2, \quad 0 \le x \le y \le 1.$. Find marginal density function of X.

Solution:

Given $f(x,y)=2$, $0 \le x \le y \le 1$

To find marginal density function of x:

$$g(x)=\int_{-\infty}^{\infty} f(x,y)dy = \int_{x}^{1} 2dy = 2[1-x], \quad 0 \le x \le y.$$

Example:7

If the joint probability density function of X and Y is given by

$$f(x,y)=\begin{cases}\frac{1}{8}(6-x-y), & 0<x<2,\ 2<y<4\\ 0, & \text{otherwise}\end{cases}$$

Find (i) $P(X<1\cap Y<3)$ (ii) $P\left(X<1/Y<3\right)$ (iii) $f\left(\frac{y}{x}\right)$.

Solution:

Given $f(x,y)=\begin{cases}\frac{1}{8}(6-x-y), & 0<x<2,\ 2<y<4\\ 0, & \text{otherwise}\end{cases}$

i) To find $P(X<1\cap Y<3)$:

$$P(X<1\cap Y<3)=\int_0^1\int_2^3 f(x,y)dydx$$

$$=\frac{1}{8}\int_0^1\int_2^3(6-x-y)dydx$$

$$=\frac{3}{8}$$

ii) To find $P\left(X<1/Y<3\right)$

$$P\left(X<1/Y<3\right)=\frac{P(X<1\quad Y<3)}{P(Y<3)} \qquad \text{.....(1)}$$

To find $P(Y<3)$:

$$P(Y<3)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx$$

$$=\int_0^2\int_2^3\frac{1}{8}(6-x-y)dydx$$

$$=\frac{5}{8}$$

Equation (1) becomes $P\left(X<1/Y<3\right)=\frac{3}{5}$

iii) To find $f(y/x)$:

We know that $f(y/x) = \dfrac{f(x,y)}{f_x(x)}$

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{8}\int_{2}^{4}(6-x-y)dy$$

$$= \tfrac{1}{4}(3-x), 0 < x < 2.$$

$$f(y/x) = \frac{\frac{1}{8}(6-x-y)}{\frac{1}{4}(3-x)} = \frac{6-x-y}{2(3-x)}, \qquad 0 < x < 2, \quad 2 < y < 4.$$

Example : 8

If the joint distribution of X and Y is given by

$F(x,y) = (1-e^{-x})(1-e^{-y})$, for $x > 0, y > 0$

$= 0$, otherwise

(i) Find the marginal densities of X and Y (ii) Are X and Y independent?

(iii) $P(1 < X < 3, 1 < Y < 2)$

Solution:

Given $F(x,y) = (1-e^{-x})(1-e^{-y})$

$$= 1 - e^{-x} - e^{-y} + e^{-(x+y)}$$

The joint pdf is given by $f(x,y) = \dfrac{\partial^2 F(x,y)}{\partial x \partial y}$

$$f(x,y) = \frac{\partial^2}{\partial x \partial y}(1 - e^{-x} - e^{-y} + e^{-(x+y)})$$

$$= e^{-(x+y)}$$

$$f(x,y) = e^{-(x+y)}, x \geq 0, y \geq 0$$

i) The marginal density function of X is $f(x) = \int_{-\infty}^{\infty} f(x,y)dy$

$$f(x) = \int_{0}^{\infty} e^{-(x+y)}dy = e^{-x}, x \geq 0$$

The marginal density function of Y is $f(y) = \int_{-\infty}^{\infty} f(x,y)dx$

$$f(y) = \int_0^\infty e^{-(x+y)} dx = e^{-y}, y \geq 0$$

ii) Consider $f(x).f(y) = e^{-x}e^{-y} = e^{-(x+y)} = f(x,y)$

ie) X and Y are independent.

iii) $P(1 < X < 3, 1 < Y < 2) = P(1 < X < 3).P(1 < Y < 2)$

$$= \int_1^3 f(x)dx . \int_1^2 f(y)dy = \int_1^3 e^{-x}dx \int_1^2 e^{-y}dy$$

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

Exercises:

1. The joint p.d.f. of the two dimensional random variable is,

$$f(x,y) = \begin{cases} \frac{8xy}{9}, & 1 < x < y < 2 \\ 0, & \text{otherwise} \end{cases}$$

(i) Find the marginal density functions of X and Y.

(ii) Find the conditional density function of Y given X=x.

2. If the joint Probability density function of two dimensional R.V (X,Y) is given by

$$f(x,y) = \begin{cases} x^2 + \frac{xy}{3}, & 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Show that X and Y are not independent.

Covariance

It is useful to measure of the relationship between two random variables is called covariance. To define the covariance we need to describe the expected value of a function of two random variables C(x,y).

Covariance:

If X and Y are random variables, than covariance between X and Y is defined as

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

$$= E\{XY - XE(Y) - E(x)y + E(X)E(Y)\}$$

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

Covariance (X, Y) = $E(XY) - E(X)E(Y)$ ……. (A)

If X and y are independent, then $E(XY) = E(X)E(Y)$ …… (B)

Substituting (B) in (A), we get Covariance (x, y) = 0

If X and Y are independent, then $Cov(X, Y) = 0$

Correlation:

If the change in are variable affects a change in the other variable, the variable are said to be correlated In a invariable distribution we may be interested to find out if there is any correlation or co-variance between the two variables under study.

Types of correlation:

1) Positive correlation
2) Negative Correlation

Positive Correlation:

If the two variables deviate in the same direction i.e. If the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive.

Example: The Correlation between

a) The height, and weight of a group of person and
b) Income and expenditure

Negative Correlation:

If the two variable constancy deviate in opposite directions i.e. if (increase 9or decrease) in one result in corresponding decrease (or increase) in the other correlation , is said to be negative.

Example: The Correlation between

a) Price and demand of a commodity and
b) The correlation between volume and pressure of a perfect gas.

Measurement of Correlation:

We can measure the correlation between the two variables by using Karl-Pearson's co -efficient of correction.

Karl-Pearson's Co-Efficient of Correlation:

Correlation co-efficient between two random variable X and Y usually denotes by (X,Y) is a numerical measure of linear. Karl Pearson's co-efficient of correlation between x & y is

r = $1 - 6 \sum_{i=1}^{n} d_i^2 / n(n^2 - 1)$, , where $d_i \quad x_i - y_i$ =

Relationship between them and detained as

$$r(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$ Where COV (X,Y) $= \frac{1}{n}\sum XY - \overline{XY}$

$$\sigma_X = \sqrt{\frac{1}{n}\sum X^2 - \overline{X_2},\ \overline{X} = \frac{\sum X}{n}}$$

(n $\sigma_Y = \sqrt{\frac{1}{n}\sum Y^2 - \overline{Y_2}}$ is the number of items in the given data)

Note:

1. Correlation coefficient may also be denoted by r(x,y)

2. If r(x,y) = 0, we say that x & y are uncorrelated.

3. When r = 1 , the correlation is perfect.

Example :9

Calculate the Correlation co-efficient for the following heights (in inches) of father x and their sons y.

Solution:

Method : 1

X	**Y**	**XY**	**X^2**	**Y^2**
67	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184

70 72	69 71	4836 5112	4900 5184	4761 5041
∑(x) =544	∑ (y) = 552	∑XY = 37560	∑ x^2 = 37028	∑ y^2 = 38132

Now

$\overline{X} = 544/8 = 68$

$\overline{Y} = 544/8 = 69$

$\overline{X}\ \overline{Y} = 68 * 69 = 4692$

$\sigma_X = \sqrt{1/n \sum x^2 - \overline{x^2}}$

$= \sqrt{37028/8 - 4624}$ =2.121

$= \sqrt{38132/8 - 4761}$ =2.345

r(X,Y)= $\frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

=1/n$\sum xy - \overline{x}\ \overline{y} / \sigma_x . \sigma_y$

=1/8*37560-4692/2.121*2.345

= 3/4.973

=0.6032

It is positive correlation.

Example:10 Find the co-efficient of Correlation between industrial productions and expose using the following data

Production (x)	55	56	58	59	60	60	62
Export (y)	35	38	37	39	44	43	44

Solution :

X	Y	U =X-58	V =Y-40	UV	U2	V2
55	35	-3	-5	15	9	25
56	38	-2	-2	4	4	4

58	37	0	-3	0	0	9
59	39	1	-1	-1	1	1
60	44	2	4	8	4	16
60	43	2	3	6	4	9
62	44	4	4	16	16	16
		$\sum$ U=4	$\sum$ U=0	$\sum$ UV=48	$\sum U^2 = 38$	$\sum V^2 = 8$

Now $\overline{U} = \sum U / n = 4/7$ =0.5714

$\overline{V} = \sum V / n = 0$ …… (1)

$\sigma_U = \sqrt{\sum U^2 / n - \overline{U}^2} = \sqrt{38/7 - (0.5714)^2}$ =2.2588…….. (2)

$\sigma_V = \sqrt{\sum V^2 - \overline{V}} = \sqrt{80/7 - 0} =$ 3.38……. (3)

$\therefore r = (X,Y) = r(U,V) = COV(U,V) / \sigma_U * \sigma_V = 6.857/2.258 * 3.38 = 0.898$[using (1), (2) & (3)]

r = 0.79

The value between 0 to 1. So it is positive correlation.

Example :11

Find the Correlation co-efficient for the following data.

X	10	14	18	22	26	30
Y	18	12	24	6	30	36

Solution:

X	**Y**	**U =X-22/4**	**V =Y-24/6**	**UV**	**U²**	**V²**
10	18	-3	-1	3	9	1
14	12	-2	-2	4	4	4
18	24	-1	0	0	1	0
22	6	0	-3	0	0	9
26	30	1	1	1	1	1
30	36	2	2	4	4	4

		$\sum U=-3$	$\sum V=-3$	$\sum UV=12$	$\sum U^2 = 19$	$\sum V^2 = 19$

Now $\overline{U} = \sum U / n = -3/6 = -0.5$ ……(1)

$\overline{V} = \sum V / n = -3/6 = 0.5$ ……(2)

$$COV(U,V) = \frac{\sum UV}{n\overline{UV}} = 1.75$$

$$\sigma_U = \sqrt{\sum U^2 - \overline{U}^2}$$

$= \sqrt{19/6} - (0.5)^2 = 1.708$ …… (3)

$\therefore r(x, y) = 0.6$

The value between 0 to 1. So it is positive correlation

Rank Correlation:

Let us suppose that a group of n individuals are arranged in order of merit or proficiently in possession of two characteristics A & B.

r = $1 - 6 \sum_{i=1}^{n} d_i^2 / n(n^2 - 1)$, , where d_i $x_i - y_i$ =

Note:

This formula is called a Spearman's formula .

Solved Problems on Rank Correlation:

Example :12

Find the rank correlation co-efficient from the following data:

Rank in X	1	2	3	4	5	6	7
Rank in Y	4	3	1	2	6	5	7

Solution

X	Y	$d = x_i - y_i$	di^2
1	4	-3	9

2	3	-1	1
3	1	2	4
4	2	2	4
5	6	-1	1
6	5	1	1
7	7	0	0
		$\sum d_i = 0$	$\sum d_i^2 = 20$

Rank Correlation co-efficient

$$r = 1 - 6 \sum_{i=1}^{n} d_i^2 / n(n^2 - 1), \quad \text{, where} \quad d_i \quad x_i - y_i =$$

$= 1 - 6 \times 20 / 7(49-1) = 0.6429$

Example : 13 The ranks of some 16 students in mathematics & physics are as follows. Calculate rank correlation co-efficient for proficiency in mathematics & physics.

Rank in Mathematics	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Rank in Physics	1	10	3	4	5	7	2	9	8	11	15	9	14	12	16	13

Solution:

Rank in Mathematics(X)	Rank in Physics(Y)	$d_i = X_i - Y_i$	d_i^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0

6	7	-1	1
7	2	5	25
8	9	-1	1
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
		$\sum d_i = 0$	$\sum d_i^2 = 136$

Rank correlation co-efficient

$r = 1 - 6 \sum_{i=1}^{n} d_i^2 / n(n^2 - 1)$, where $d_i = x_i - y_i$

r = 0.8

Example : 14

10 competitors in a musical test were ranked by the 3 judges X, Y, Z in the following order

	A	B	C	D	E	F	G	H	I	J
Rank in X	1	6	5	10	3	2	4	9	7	8
Y	3	5	8	4	7	10	2	1	6	9
Z	6	4	9	8	1	2	3	10	5	7

Using Rank correlation method, discuss which panel of Judges has the nearest approach to common likings of music.

X	Y	Z	$D_1 = x_i - y_i$	$D_2 = y_i - z_i$	$D_3 = x_i - z_i$	D_1^2	D_2^2	D_3^2

1	3	6	-2	-3	-5	4	9	25
6	5	4	1	1	2	1	1	4
5	8	9	-3	-1	-4	9	1	16
10	4	8	6	-4	2	36	16	4
3	7	1	-4	6	2	16	36	4
2	10	2	-8	8	0	64	64	0
4	2	3	2	-1	1	4	1	1
9	1	10	8	-9	-1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	-1	2	1	1	4	1

$\sum d_1^2 = 200$ $\sum d_2^2 = 214$ $\sum d_3^2 = 60$

The rank correlation between X & Y is

$$r^1 = 1 - 6 \sum_{i=1}^{n} d_i^2 / n(n^2 - 1) = -0.212$$

The rank correlation between Y& Z is

$$r^2 = 1 - 6 \sum_{i=1}^{n} d_i^2 / n(n^2 - 1) = -0.296$$

The rank correlation between X & Z is

$$r^3 = 1 - 6 \sum_{i=1}^{n} d_i^2 / n(n^2 - 1) = 0.636$$

Since the rank correlation between X & Z is maximum and also positive, We conclude that the pair of Judges X & Z has the nearest approach to common likings of music.

Exercises:

1) Calculate the Karl Pearson's co-efficient of correlation from the following data

X	25	26	27	30	32	35
Y	20	22	24	25	26	27

2) Find the co-efficient of correlation of the advertisement cost & sales from the following data

Cost:	39	65	62	90	82	75	98	36	78
Sales:	47	53	58	86	62	68	91	51	84

REGRESSION

Definition:

Regression is a mathematical measure of the average relationship between two or more variables in terms of the original limits of the data.

Lines of regression:

If the variables in a bivariate distribution are related we will cluster around some curve called of regression. If the curve is a straight line, it is and called the line of regression and there is said to be linear regression is said to be curve linear.

The line of regression of y on x is given by $y-\bar{y}=r.\frac{\partial y}{\partial x}(x-\bar{x})$

where r is the correlation coefficient, ∂_y and ∂_X are standard deviation.

The line of regression of X on Y is given by $x-\bar{x}=r.\frac{\partial y}{\partial x}(y-\bar{y})$

Angle between two line of regression:

If the equation of lines of regression of Y on X and X on Y are

$y-\bar{y}=r.\frac{\partial y}{\partial x}(x-\bar{x})$ and $x-\bar{x}=r.\frac{\partial x}{\partial y}(y-\bar{y})$

The angle $'\theta'$ between the two line of regression is given by

$$\tan\theta=\frac{l-r^2}{r}\left(\frac{\partial y\partial x}{\partial x^2+\partial y^2}\right)$$

Regression coefficients:

Regression coefficient of Y on X , $r\frac{\partial Y}{\partial X}=b_{YX}$ ……. (1)

Regression coefficient of X on Y , $r\frac{\partial X}{\partial Y}=b_{XY}$ …….. (2)

From (1) and (2) we get

$$r\frac{\partial Y}{\partial X}r\frac{\partial X}{\partial Y}=b_{YX}*b_{YX}$$

Correlation coefficient $r = \pm\sqrt{b_{XY} * b_{YX}}$

The regression coefficients b_{YX} and b_{yx} can be easily obtained by using the following formula.

$$b_{YX} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$b_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$$

Solved Problems on Regression:

Example:15

The equations of two regression lines are 3x+12y=19, 3y+9x=46. Obtain the mean value of X and Y.

Solution:

Given the lines are 3x+12y=19,

3y+9x=46

Since both are passing through $(\bar{x}, \bar{y})$, we get

$3\bar{x} + 12\bar{y} = 19$ ……(1)

$9\bar{x} + 3\bar{y} = 46$ ……..(2)

Solving equation (1) & (2) we get $33\bar{y} = 11$

$\bar{y} = \frac{11}{33} = 0.33$, $\bar{y}$ value sub in equation (1) we get $\bar{x} = 5$

$(\bar{x}, \bar{y}) = (5, 0.33)$

Example:16

From the following data, find

i)The two regression equations.

ii) The co-efficient of correlation between the marks in economics and statistics.

iii)The most likely marks in statistics when marks in economics are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Solution:

X	Y	$X-\overline{X}=X-32$	$Y-\overline{Y}=Y-38$	$(X-\overline{X})^2$	$(Y-\overline{Y})^2$	$(X-\overline{X})(Y-\overline{Y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
$\Sigma X=320$	$\Sigma Y=380$	$\Sigma(X-\overline{X})=0$	$\Sigma(Y-\overline{Y})=0$	$\Sigma(X-\overline{X})^2=140$	$\Sigma(Y-\overline{Y})^2=398$	$\Sigma(X-\overline{X})(Y-\overline{Y})=-93$

Here $\overline{X}=\frac{\sum X}{n}$ and $\overline{Y}=\frac{\sum y}{n}$

$=\frac{320}{10}=32$ $=\frac{380}{10}=38$

Coefficient of regression of Y on X is

$$b_{YX} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} = \frac{-93}{140}$$

Coefficient of regression of X on Y is

$$b_{XY} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (Y-\overline{Y})} = \frac{-93}{398} = -0.2337$$

Equation of the line of regression of X on Y is

$$x-\overline{x} = b_{XY}(y-\overline{y})$$

X – 32 =0.2337(Y-38)

X= -0.2337 y+0.2337 *38 +32

X=-0.23374 +40.8806

Equation of the line of regression of Y on X is

$$y-\overline{y} = b_{yx}(x-\overline{x})$$

Y-38 = -0.6643(x-32)

Y = - 0.6643 x+38+0.6643*32 =-0.6642x+59.2576

Now we have to find the most likely marks in statistics (Y) when marks in economics (X) are 30.we use the line of regression of Y on X.

Y = -0.6643x+59.2575

Put x=30, we get

Y = -0.6643*30+59.2536 =39.3286 =39

Example :17

Height of father and sons are given in centimeters

X:Height of father	150	152	155	157	160	161	164	166
Y:Height of son	154	156	158	159	160	162	161	164

Find the two lines of regression and calculate the expected average height of the son when the height of the father is 154 cm.

Solution:

Let 160 and 159 be assured means of x and y.

x	y	U=X-160	V=Y-159	u^2	v^2	uv
150	154	-10	-5	100	25	50
152	156	-8	-3	64	9	24
155	158	-5	-1	25	1	5
157	159	-3	0	9	0	0
160	160	0	1	0	1	0
161	162	1	3	1	9	3
164	161	4	2	16	4	8
166	164	6	5	36	25	30
		$\sum U = -15$	$\sum V = 2$	$\sum U^2 = -15$	$\sum V^2 = 74$	$\sum UV = 120$

Now $\overline{X} = 158.13$ and $\overline{Y} = 159.25$

Since regression coefficient are independent of change and of origin we have regression coefficient of Y on X

Coefficient of regression of Y on X is

$$b_{YX} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} = \frac{990}{1783} = 0.555$$

Coefficient of regression of X on Y is

$$b_{XY} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (Y-\overline{Y})} = 1.68.$$

Exercise:

1. The two lines of regression are $8x - 10y = 66$ *and* $40x - 18y - 214 = 0$. The variance of X is 9. Find i) the mean values of X and Y ii) Correlation between X and Y.

S.NO	Questions	OPT 1	OPT 2	OPT3	OPT 4
1	The correlation coefficient is used to determine_____	A specific value of the y-variable given a specific value of the x-variable	A specific value of the x-variable given a specific value of the y-variable	The strength of the relationship between the x and y variables	none of these
2	If X and Y are independent , then	E(XY)=0	E(X) E(Y)=0	Cov(X,Y) =0	E(XY)=1
3	The relationship between three or more variables is studied with the help of _____ correlation.	multiple	rank	perferct	spearman's rank
4	The coefficient of correlation is under-root of two _____	regression coefficients	rank coefficient	Regression equation	regression line
5	The coefficient of correlation___	has no limits	can be less than 1	can be more than 1	varies between + or - one
6	which of the following is the highest range of r_____	0 and 1	minus one and 0	minus one and one	zero
7	The coefficient of correlation is independent of _____	change of scale only	change of origin only	both change of scale and origin	change of variables
8	The coefficient of correlation___	cannot be positive	cannot be negative	can be either positive or negative	zero
9	COV(X,Y)= _____	E(XY)-E(X) E(Y)	E(XY)+E (X)E(Y)	E(XY)	Var(X,Y)
10	Two random variables with non zero correlation are said to be___	correlation	regression	rank	variables
11	Correlation means relationship between _____ variables	two	one	two or more	three
12	Two random variables X and Y with joint pdf f(x,y) is said to independent if ____	f(x,y) = f(x) +f(y)	f(x,y) = f(x) / f(y)	f(x,y) = f(x) * f(y)	f(x,y) = f(x) - f(y)
13	The covariance of two independent random variable is _____	Zero	two	three	two or more
14	Two random variables are said to be orthogonal if _____	correlation is zero	rank is zero	covariance is zero	one

OPT 5	**OPT 6**	**ANSWERS**
		The strength of the relationship between the x and y variables
		Cov(X,Y) =0
		multiple
		regression coefficient
		varies between + or - one
		minus one and one
		both change of scale and origin
		can be either positive or negative
		E(XY)-E(X)E(Y)
		regression
		two or more
		f(x,y) = f(x) * f(y)
		Zero
		correlation is zero

15	Two random variables are said to be uncorrelated if correlation coefficient is____	zero	one	two or more	orthogonal
16	Regression analysis is a mathematical measures of the average relationship between _____variable	two or more	one	Two variables	three
17	The regresision analysis confined to ther study of only two variable at a time is called_____regression	Simple	Multiple	Linear	two
18	If r=0, then the regression coefficient are_____	zero	one	threee	constant
19	The equation of the fitted stright line is _____	y=ax+b	y=a+bx	y=mx+c	y=mx
20	If X=Y , then correlation cofficient between them is _____	1	zero	less than one	gerater than one
21	The greater the value of r_____obtained through regression analysis	the better are estimates	the worst are the estimates	really makes no difference	good estimates
22	Where r is zero the regression lines cut each other making an angle of_____	30 degree	60 degree	90 degree	neither of the above
23	The father the two regression lines cut each other_____	Greater will be degree of correlation	The less will be the degree of correlation	does not matter	the worst are the estimates
24	The regression lines cut each other at the point of :_____	Average of X and Y	Average of X only	Average of Y only	average of both (a) and (b)
25	When the two regression lines coincide, then r is :_____	0	-1	1	0.5
26	The variable , we are trying to predict is called the _____	depentent variable	indepent variable	constant	normal
27	Both the regression coefficients cannot _____one	exceed	exact	plus or minu	negative
28	The regression analysis measures _____between variables	dependence	independenc e	constant	normal
29	If the possible values of (X,Y) are finite, then (X,Y) is called a_____	two dimensional random variable	onedimension al random variable	both a and b	infinte

		zero
		two or more
		Simple
		zero
		y=ax+b
		1
		the better are estimates
		neither of the above
		the less will be the degree of correlation
		average of X and Y
		1
		dependent variable
		exceed
		dependence
		two dimensional random variable

30	If X & Y are continuous random variable , then f(x,y) is _____	joint probability function	joint probability density function	both a and b	infinte
31	Joint probability is the probability of the _____occurrence of two or more events.	Simultaneous (or) joint	Conditional	Mariginal probability	density function
32	The order of arrangement is important in _____	permutation	Gambling	joint	density
33	If X & Y are____random variable , then f(x,y) is called joint probability function.	discrete	continuous	both a and b	infinte
34	If the value of y decreases as the value of x increases then there is _____correlation between two variables.	negative	perfect positive	both a and b	infinte
35	The correlation between the income and expenditure is _____	positive	negative	finite	both a and b
36	correlation between price and demand of commodity is ____	positive	finite	negative	both a and b
37	If X and Y are independent , then _____	E(XY) = E (X) + E(Y)	E(XY) = E (X) - E(Y)	E(XY) = E (X) E(Y)	E(XY) = E(X)/E (Y)
38	correlation coefficient does not exceed _____	unity	5	0	2
39	Two independent variables are _____	correlated	uncorrelated	both a and b	positive
40	In Rank correlation the correction factor is added for each _____value.	repeated	Non-repeated	indefinite	both a and b
41	When r = 1 or -1 the the line of regression are ____ to each other.	parallel	perpendicular	straight line	circular
42	If the curve is a straight line, then it is called the ____	the line of correlation	the line of regression	covariance	both a and b
43	If the curve is not a straight line, then it is called the ____	covariance	the line of correlation	the curvilinear	the line of regression
44	when r is ____ the correlation is perfect and positive.	1	2	3	0
45	The coefficient of correlation is independent of change of_____ and _____	scale,origin	vector,origin	variable, constant	interer, origin
46	When r = 0 the line of regression are _____to each other.	parallel	perpendicular	straight line	circular

		both a and b
		Simultaneous (or) joint
		permutation
		continuous
		negative
		positive
		negative
		E(XY) = E(X) E(Y)
		unity
		uncorrelated
		repeated
		parallel
		the line of regression
		the curvilinear
		1
		scale,origin
		perpendicular

47	A Mathematical measure of the average relationship between two variables is called_____	correlation	regression	rank	variables
48	Cov(X,Y)=_____	E[{ X- E(X) }* { Y – E (Y) }]	E[{ X- E(X) }+ { Y – E (Y) }]	E[{ X- E(X) }- { Y – E (Y) }]	E[{ X- E(X) } { Y – E(Y) }]
49	The coefficient of correlation_____	is the square of the coefficient of determination	is the square root of the coefficient of determination	is the same as r-square	can never be negative
50	The correlation between two variables is of order_____	2	1	0	3

		correlation
		E[{ X- E(X) } { Y – E(Y) }]
		is the square root of the coefficient of determination
		0

UNIT III

Testing of Hypothesis

Introduction:

Many problems in engineering require that we decide whether to accept or reject a statement about some parameter. The statement is called a hypothesis and the decision making procedure about the hypothesis is called hypothesis testing.

Population:

A population in statistics means a set of objects or mainly the set of numbers which are measurements or observations pertaining to the objects.

Sampling:

A part selected from the population is called a sample. The process of selection of a sample is called sampling.

Sampling Distribution:

The sample mean, the sample median and the sample standard deviation are examples of random variables whose values will vary from sample to sample. Their distributions, which reflect such chance variations, play an important role in statistics and they are referred to as sampling distributions.

If we draw a sample of size n from a given finite population of size N then the total number of possible samples is NC_n

$$NC_n = \frac{N!}{n!(N-n)!} = k$$

For each of these K samples we can compute some statistics say $t = t(x_1, x_2, x_3 ...., x_n)$ in particular the mean $\bar{x}$, variance (s^2) etc. The set of the values of the statistic so obtained, one for each sample constitutes the sampling distribution of the statistic.

Standard Error:

The standard deviation of sampling distribution of a statistic is known as standard error and it is denoted by (S.E)

Testing a hypothesis:

On the basis of sample information, we make certain decisions about the population . In taking such decisions we make certain assumptions. These assumptions are known as statistical hypothesis are tested.

Assuming the hypothesis correct we calculate the probability of getting the observed sample. If this probability is less than a certain assigned value, the hypothesis is to be rejected.

Null hypothesis:

Null hypothesis is based for analyzing the problem . Null hypothesis is the hypothesis of np difference. It is denoted by H_0. It is defined as a definite statement about the population parameter.

Alternative hypothesis:

Any hypothesis which is complementary to the null hypothesis is called alternative hypothesis . It is denoted by H_1.

Rule:

1.If we want to test the significance of the difference between a statistic and the parameter or between two sample statistics then we setup the null hypothesis that the difference is not significant.

2.If we want to test any statement about the population , we set up the null hypothesis that it is true.

Types of errors:

Type 1 Error: Reject H_0 when it is true.

Type 2 error: Accept H_0 when it s wrong.

P (Reject H_0 when it is true) = P (Type I error) = α

P (Accept H_0 when it is wrong) = P (Type II error) = β

The sizes of the type I and type II errors are also known as producer's risk and consumer's risk respectively.

α = P (Rejecting a good lot)

β = P (Accepting a bad lot)

Level of significance:

The probability that the value of static lies in the critical region is called as level of significance.

Test of significance of small samples:

When the size of the sample (n) is less than 30, then that sample is called a small sample. The following are some important test for small samples.

(i) Student's 't' test

(ii) F- Test

Test for single mean (Student's 't' test):

Suppose we want to test

(a) If a random sample x_i of size n has been drawn from a normal population with a specified mean μ_0

(b) If the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

In this case, the statistic is given by,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ follows students t- distribution with (n-1) degrees of freedom. We now compare the calculated value of t with tabulated value at a certain level of significance. If calculated $|t|$ > tabulated t, null hypothesis is rejected and if calculated $|t|$ < tabulated t, null hypothesis may be accepted.

Assumption for students t – test:

1. The parent population from which the sample is drawn is normal.
2. The sample observations are independent

The population standard deviation is unknown.

Example : 1

A machinist is making engine parts with axle diameters of 0.700 inch. A random sample of ten parts shows a mean diameter of 0.742 inch with a standard deviation of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications

Solution:

Given, $\mu = 0.700$ inches

$\bar{x} = 0.742$ inches

$s = 0.040$ inches and $n = 10$.

$H_0 : \mu = 0.700$ i.e. the product is conforming to specifications.

$H_1 : \mu \neq 0.700$

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2}/\sqrt{n}} = \frac{\bar{x} - \mu}{\sqrt{s^2}/\sqrt{n-1}} \sim t_{(n-1)}$$

$$t = \frac{0.742 - 0.700}{\sqrt{(0.040)^2}/\sqrt{10-1}} = 3.15$$

t follows student t distribution.

The table value of t at 5% level of significance for 9 degrees of freedom is $t_{0.05} = 1.833$

Calculated t > tabulated t, H_0 rejected.

Example :2

A random sample of 10 boys had the following IQs. 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean IQ of 100? Find a reasonable range in which most of the mean IQ values of samples of 10 boys lie.

Solution

H_0 : The data support the assumption of a population mean IQ of 100 in the population

$H_0 : \mu = 100$

$H_1 : \mu \neq 100$

Here n = 10, $\bar{x} = \frac{972}{10} = 97.2$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 203.73$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{197.2 - 1001}{14.27/\sqrt{10}} = 0.62$$

Tabulated value of t_0 for (10-1) degrees of freedom is 2.26

$t < t_0$ H_0 may be accepted and we conclude that the data are consistent.

The 95% confidence limits are given by $\bar{x} \pm t_{0.05}\frac{s}{\sqrt{n}} = 97.2 \pm 2.26(4.514)$

= 107.40 and 87.00

Exercise

1. A sample of 26 bulbs gives a mean life of 990 hours with a SD of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not up to the standard?

2. The following data gives the length of 12 samples of Egyptian cotton taken from a large consignment 48,46,49,46,52,45,43,47,47,46,47,50. Test if the mean length of the consignment be taken as 46.

Students 't' test for difference of Means: (Corrected t test or Paired t test)

To test the significant difference between two means $\overline{x_1}$ and $\overline{x_2}$ of samples of sizes n_1 and n_2 the statistic is

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{s^2 \frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n} x_i \quad , \quad \bar{y} = \frac{1}{n_2} \sum_{j=1}^{n} y_j \quad ,$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{j=1}^{n} (y_i - \bar{y})^2 \right]$$

Degrees of freedom = $n_1 + n_2 - 2$

Example: 3

Below are the gain in weight in lbs of pigs fed on the 2 diets A and B.

Gain in weight

Diet A: 25,63,30,34,24,14,32,24,30,31,35,25

Diet B: 44,34,22,10,47,31,40,30,32,35,80,21,35,29,22

Test if the two diets differ significantly as regards to their effect on increasing the weight.

Solution:

$H_0 : \mu_x = \mu_y$ i.e. There is no significant difference between the mean increase in the weights due to diet A and B.

$H_1 : \mu_x \neq \mu_y$

Diet A $\sum x = 336$ $\sum (x-\bar{x})^2 = 380$ $n_1 = 12$ $n_2 = 15$

Diet B $\sum y = 450$ $\sum (y-\bar{y})^2 = 1410$ $\bar{x} = 28$ $\bar{y} = 30$ $S^2 = 171.6$

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

$$= \frac{30 - 28}{\sqrt{71.6\left(\frac{1}{12} + \frac{1}{15}\right)}} = 0.609$$

t_0 at 5% level of significance for (12+15-2)25 degrees of freedom is 2.06

$t < t_0 \quad \therefore \quad H_0$ may be accepted and we may conclude that the two diets do not differ significantly.

Example : 4

The student of 6 randomly chosen sailors are in inches: 63,65,68,69,71,72

These of 10 randomly chosen sailors are (in inches): 61,62,65,66,69,69,70,71,72,73

Discuss the height that these data throw on the suggestions that the sailors are on the average taller than soldiers.

Solution:

Given $n_1 = 6$, $n_2 = 10$

$\bar{x} = 68$ $\bar{y} = 67.8$

x	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
63	-5	25	61	-6.8	46.24
65	-3	9	62	-5.8	33.64
68	0	0	65	-2.8	7.84
69	1	1	66	-1.8	3.24
71	3	9	69	1.2	1.44
72	4	16	69	1.2	1.44
Total		60	70	2.2	4.84
			71	3.2	10.24
			72	4.2	17.64
			73	5.2	27.04
			Total		153.6

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + \sum_{j=1}^{n}\left(y_i - \bar{y}\right)^2\right]$$

$= 15.2571.$

$H_0 : \mu_1 = \mu_2$

Test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{68 - 67.8}{\sqrt{15.271\left(\frac{1}{6} + \frac{1}{10}\right)}} = 0.099$$

Reject H_0 if $|t| < 1.76$, We accept H_0 at 5% level of significance.

$\therefore t = 0.099$

Exercise:

1. Average number of articles produced by two machines per day is 200 and 250 with its standards deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machine equally efficient at 1% level of significance?

2. Two horses A and B were tested according to the time (in seconds) to run a particular race with following results:

   Horse A: 28 30 32 33 33 29 34

   Horse B: 29 30 30 24 27 29

F -distribution (Test for variance) – Snedecor's F – distribution:

To test whether if there is any significant difference between two estimates of population variance. To test if the two samples have come from the same population, we use F test.

$H_0 : \sigma_1^2 = \sigma_2^2$ (i.e.) Population variations are same.

Under H_0 the test statistic is $F = \frac{S_x^2}{S_y^2}$,

Where $S_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^2$ & $S_y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(y_j - \bar{y}\right)^2$ and it follows F distribution with (v_1, v_2) degrees of freedom where $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$.

Note:

1. We will take greater of the variance S_1^2 or S_2^2 in the numerator and adjust for the degrees of freedom accordingly

$$F = \frac{Greater\ \text{var}iance}{Smaller\ \text{var}iance}$$

2. If sample variance s^2 is given we can obtain population variance S^2 by using the relation $ns^2 = (n-1)S^2$

Example :5

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 another sample of 10 observations it was 102.6. Test whether this difference is significant at 5% level.

Solution:

Given $n_1 = 8$ $n_2 = 10$

$\sum\left(x - \bar{x}\right)^2 = 84.4$ $\sum\left(y - \bar{y}\right)^2 = 102.6$

$S_x^2 = \frac{1}{n_1 - 1}\sum\left(x - \bar{x}\right)^2 = 12.057$

$S_y^2 = \frac{1}{n_2 - 1}\sum\left(y - \bar{y}\right)^2 = 11.4$

Steps:

1. The parameter of interest is σ_x^2 & σ_y^2

2. $H_0 : \sigma_x^2 = \sigma_y^2 = \sigma^2$

3. $H_1 : \sigma_x^2 = \neq \sigma_y^2$

4. $\alpha = 0.05$, d.f $(V_1) = n_1 - 1 = 7$

d.f $(V_2) = n_2 - 1 = 9$

5. $F = \frac{S_x^{\,2}}{S_y^{\,2}} = 1.057$

6. Reject H_0 if F>3.29 (from table F)

7. Computations: $F = \frac{12.057}{11.42} = 1.057$

8. Conclusion: Since Tabulated $F_{0.05}$ for (7,9) degrees of freedom is 3.29 $F < F_0$

$\therefore$ H_0 is accepted.

Example :6

Two random samples gave the following results.

Sample	Size	Sample mean	Sum of the squares of deviations from the mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population.

Solution:

To test if two independent samples have been drawn from the same normal population, we have to test

1. The equality of population means and
2. The equality of population variances

$H_0 : \mu_1 = \mu_2 \;\&\; H_0 : \sigma_1^{\,2} = \sigma_2^{\,2}$

$n_1 = 10 \qquad n_2 = 12 \qquad \overline{x_1} = 15 \qquad \overline{x_2} = 14$

$\sum (x_1 - \overline{x_1})^2 = 90 \qquad \sum (x_2 - \overline{x_2})^2 = 108$

F Test

$S_{x_1}^{\,2} = \frac{1}{n_1 - 1} \sum (x_1 - \overline{x_1})^2 = \frac{90}{9} = 10$

$$S_{x_2}{}^2 = \frac{1}{n_2 - 1}\sum\left(x_2 - \overline{x_2}\right)^2 = \frac{108}{11} = 9.827$$

$$F = \frac{S_1^{\,2}}{S_2^{\,2}} = 1.078$$

Tabulated $F_{0.05}$ for (9,11) degrees of freedom is 2.90 $F < F_0$

$\therefore$ H_0 is accepted.

t Test:

$H_0 : \mu_1 = \mu_2$

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum\left(x_1 - \overline{x_1}\right)^2 + \sum\left(x_2 - \overline{x_2}\right)^2\right] = 9.9$$

$\therefore$ t = 0.742

$t_{0.05}$ for 20 degrees of freedom is 2.086

$|t| < t_0$ $\therefore$ H_0 is accepted.

Both the hypothesis are accepted.

$\therefore$ We may consider that the given samples have been drawn from the same population.

Example:7

A group of 10 rats fed on diet A and another group of 8 rats fed on diet B, recorded the following increase in weight.

Diet A	5	6	8	1	12	4	3	9	6	10
Diet B	2	3	6	8	10	1	2	8		

Find if the variances are significantly different.

Solution:

Given $n_1 = 10$ $n_2 = 8$

Diet-A			Diet-B		
x	$(x - \overline{x})$	$(x - \overline{x})^2$	y	$(y - \overline{y})$	$(y - \overline{y})^2$

5	-1.4	1.96	2	-3	9
6	-0.4	0.16	3	-2	4
8	1.6	2.56	6	1	1
1	-5.4	29.16	8	3	9
12	5.6	31.36	1	-4	16
4	-2.4	5.76	10	5	25
3	-3.4	11.56	2	-3	9
9	2.6	6.76	8		
6	-0.4	0.16	$\sum y = 40$	Total = 82	
10	3.6	12.96			
$\sum x = 64$	Total = 102.4				

$\bar{x} = 6.4$ and $\bar{y} = 5$

$$\sum (x - \bar{x})^2 = 102.4 \qquad \sum (y - \bar{y})^2 = 82$$

$$S_x^{\,2} = \frac{1}{n_1 - 1} \sum (x - \bar{x})^2 = 11.378$$

$$S_y^{\,2} = \frac{1}{n_2 - 1} \sum (y - \bar{y})^2 = 11.71$$

Steps:

1. The parameter of interest is $\sigma_x^{\,2} \,\&\, \sigma_y^{\,2}$

2. $H_0 : \sigma_x^{\,2} = \sigma_y^{\,2} = \sigma^2$

3. $H_1 : \sigma_x^{\,2} = \neq \sigma_y^{\,2}$

4. $\alpha = 0.05$, $d.f(V_1) = n_1 - 1 = 9$, $d.f(V_2) = n_2 - 1 = 7$

5. $F = \dfrac{S_x^{\,2}}{S_y^{\,2}} = 1.02$

6. Reject H_0 if F > 5.19 (from table F)

7. Computations: 1.02 < 5.19

8. Conclusion: Since Tabulated $F_{0.05}$ for (9,7) degrees of freedom is 5.19 $F < F_0$

$\therefore$ H_0 is accepted. We conclude that the difference is not significant.

Exercise:

The nicotine content in milligram of two samples of tobacco were found to be as follows.

Sample A: 24, 27, 26, 21, 25

Sample B: 27, 30, 28, 31, 22, 36

Can it be said that two samples come from the same normal population?

Test of significance for single mean (Normal):

Large samples:

If the size of the sample n>30, then that sample is called large sample. There are 4 important test to test the significance of large samples.

1. Test of significance for single proportion
2. Test of significance for test of difference of proportions
3. Test of significance for single mean
4. Test of significance for difference of means.

Test of significance for single mean:

Suppose we want to test whether the given sample of size n has been drawn form a population with mean μ. We set up a null hypothesis that there is no difference between $\bar{x}$ and μ where $\bar{x}$ is the sample mean.

The test statistic is $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ where σ is the standard deviation of the population. If the population standard deviation is not known use the sample standard deviation $z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$

Note:

The values $\bar{x} \pm 1.96\sigma / \sqrt{n}$ is called 95% fiducial limits or confidence limits and similarly $\bar{x} \pm 2.58\sigma / \sqrt{n}$ is called 99% confidence limits.

Example :8

A sample of 900 members has a mean 3.4 cm. and SD 2.61 cms. Is the same from a large population of means 3.25 cms and SD 2.61 cms. If the population is normal and its mean is unknown find the 95% and 98% fiducial limits of the true mean.

Solution:

Given n=900, $\mu = 3.25 cms$, $\sigma = 2.61$, $\bar{x} = 3.4 cms$

1. The parameter of interest is μ

2. H_0: The sample has been drawn from the population with mean $\mu = 3.2 cms$ and standard deviation $\sigma = 2.61 cms$

3. $H_1 : \mu \neq 3.25$

4. $\alpha = 0.05$

5. Test Statistic : $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$ n is large

6. Reject H_0 $|z| > 1.96$ $\therefore z = 1.73$

Since $|z| < 1.96$, we accept H_0 at 5% level of significance. Therefore the sample has been drawn from the large population with mean $\mu = 3.25 cms$

95% fiducial limits for the population mean μ are $\bar{x} \pm 1.96\sigma / \sqrt{n} = 3.5705 \& 0.1705$

98% fiducial limits for the population mean μ are $\bar{x} \pm 2.33\sigma / \sqrt{n} = 3.40 \pm 2.33(2.61/\sqrt{900})$

Exercise:

The average marks in Mathematics of a sample of 100 students was 51 with a S.D of 6 marks. Could this have been a random sample from a population with average marks 50.

Test of significance for difference of mean:

Let $\bar{x_1}$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2. Let $\bar{x_2}$ be the mean of a sample of size n_2 from a population with mean μ_2 and variance σ_2^2.

$H_0 : \mu_1 = \mu_2$

Test Statistic : $$z = \frac{\bar{x}_1 - \bar{x_2}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Note :

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then under $H_0 : \mu_1 = \mu_2$,

$$z = \frac{\bar{x}_1 - \bar{x_2}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1)$$

Example :9

The means of two single large samples of 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 inches.

Solution:

Given $n_1 = 1000$ $n_2 = 2000$ $\bar{x}_1 = 67.5$ $\bar{x}_2 = 68$

$H_0 : \mu_1 = \mu_2$ and $\sigma = 2.5$ inches.

$H_1 : \mu_1 \neq \mu_2$

Test Statistic : $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$= -5.1$

$|z| > 3$ H_0 is rejected.

Example:10

In a survey of buying habits 400 women shoppers are chosen at random in supermarket 'A' located in a certain section of the city. Their average weekly food expenditure is Rs.250 with a standards deviation of Rs.40. For 400 women shoppers chosen at random in supermarket 'B', the average weekly food expenditure is Rs.220 with a SD of Rs.55. Test at 1% level of significance whether the average weekly food expenditures of the two populations of shoppers are equal.

Sol: Given $n_1 = 400$, $n_2 = 400$, $\bar{x}_1 = 250$, $\bar{x}_2 = 220$, $\sigma_1 = 40$, $\sigma_2 = 55$

1. The parameter of interest is μ_1 and μ_2

2. $H_0 : \mu_1 = \mu_2$

3. $H_1 : \mu_1 \neq \mu_2$

4. $\alpha = 0.01$

5. Test Statistic : $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$= \frac{30}{\sqrt{11.5625}} = \frac{30}{3.4} = 8.8235$

$|z| = 8.8235 > 2.58$. H_0 is rejected.

Exercise:

A simple sample of heights of 6400 Englishmen has a mass of 6785 inches and a S.D of 2.56 inches, while a simple sample of heights of 1600 Australians has a mean of 68.55 inches and a S.D of 2.52 inches. Do the data indicate that Australians are on the average taller than Englishmen.

Test of significance for single proportion:

If X is the number of success in n independent trials with constant probability P of successes for each trial,

$E(x) = nP \& V(x) = nPQ$ where $Q = 1 - P$ is the probability of failure. It has been proved that for large n the Binomial distribution tends to normal distribution. Hence for large n $X \sim N(nP, nPQ)$ i.e.,

$$z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{X - nPQ} \sim N(0,1)$$

and we can apply the normal test.

Example:11

In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?

Solution:

Given n =1000, X=540

p = sample proportion of rice eaters $= \frac{540}{1000} = 0.54 = x$

P = Population proportion of rice eaters = ½ = 0.5

Q = 0.5

1. The parameter of interest is P

2. $H_0 : P = 0.5$, Both rice and wheat eater are equally popular in the state

3. $H_1 : P \neq 0.5$

4. $\alpha = 0.01$

5. Test Statistic: $$z = \frac{x - P}{\sqrt{PQ/n}} = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{1000}}} = 2.532$$

6. Conclusion: Since $z_{0.01} = 2.58$ $|z| < z_{0.01}$ $\therefore H_0$ is accepted at 1% level of significance. We conclude that rice and wheat eaters are equally popular in Karnataka

Example:12

In a study designed to investigate whether certain detonators used with explosives in a coal mining meet the requirement that at least 90% will ignite the explosives when charged it is found that 174 of f 200 detonators function properly. Test the null hypothesis P = 0.90 against the alternative hypothesis P< 0.90 at the 0.05 level of significance

Solution:

$H_0 : P = 0.90$

$H_1 : P < 0.90$

X = 174 n = 200 P = 0.90 Q = 0.10

$$z = \frac{X - nP}{\sqrt{nPQ}} = \frac{174 - 200(0.90)}{\sqrt{200 \times 0.90 \times 0.10}} = -1.41$$

Tabulated value of Z at 5% level of significance for right tail test is 1.645

$|z| < 1.645$ $\therefore H_0$ is accepted.

Exercise:

Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease is 85% in favour of the hypothesis that its more at 5% level.

Test of significance for difference of proportions:

Suppose we want to compare two distinct population with respect to the preval of a certain attribute say A, among their members. Let X1, X2 be number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two population respectively. Then the sample proportions are given by

$$P_1 = \frac{X_1}{n_1} \& P_2 = \frac{X_2}{n_2}$$

If $P_1 \& P_2$ are populations, then $E(P_1) = P_1 \& E(P_2) = P_2$

$$V(P_1)=\frac{P_1Q_1}{n_1}\ \&\ V(P_2)=\frac{P_2Q_2}{n_2}$$

Under $H_0: P_1=P_2$ the test statistic for the difference of proportions is

$$z=\frac{\overline{P_1}\ -\overline{P_2}}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim N(0,1)$$

Example:13

Random samples of 400 men and 600 women were asked whether they would like to have a fly over near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportion of men and women in favour of the proposal are same against that they are not at 5% level.

Solution:

Given $n_1=400$, $X_1=$ No of men favouring the proposal = 200

$n_2=600$, $X_2=325$

$P_1=\frac{X_1}{n_1}=\frac{200}{400}=0.5$ Similarly, $P_2=\frac{X_2}{n_2}=\frac{325}{600}=0.54$

1. The parameter of interest is $P_1\ \&\ P_2$ the difference

2. $H_0: P_1=P_2=P$

3. $H_1: P_1\neq P_2$

4. $\alpha=0.05$

5. The test statistic is

$$\therefore z=\frac{P_1-P_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim N(0,1)$$

$$\hat{P}=\frac{n_1P_1+n_2P_2}{n_1+n_2}=\frac{200+325}{400+600}=0.525$$

$$\hat{Q}=1-\hat{P}=1-0.525=0.475$$

$$\therefore z = \frac{0.5 - 0.541}{\sqrt{0.525 \times 0.475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.24$$

$|z| = 1.24$ 6. Reject H_0 if $|z| > 1.96$

7. Conclusion: $|z| < 1.96 \therefore H_0$ is accepted.

Exercise:

Before an increase in excise duty on tea, 800 persons out of a sample of 1000 persons found to be tea drinkers. After an increase in duty 800 people were tea drinkers in sample of 1200 people. Using standard error of proportion state whether there is a significant decrease in the consumption in tea after the increase in excise duty.

Chi –square distribution:

(i) Chi – Square Test of Goodness of Fit:

A very powerful test for testing the significance of the difference between theory and experiment was given by Karl pearson is 1990 and is known as " Chi square test of goodness of fit".

If $O_i (i = 1,2,...,n)$ is a set of observed or experimental frequencies and $E_i (i = 1,2,...,n)$ is the corresponding set of expected frequencies, are significant or not.

Then Karl Pearson's χ^2 is given by,

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

χ^2 is used to test whether the differences between observed and expected frequencies are significant. Degrees of freedom for B.D= n-1.

Applications of χ^2 distribution:

1. To test the goodness of fit.

2.To test the independence of attributes.

3. To test if the hypothetical value of the population vaiance is σ^2

4. To test the homogeneity of independent estimates of the population variance.

5. . To test the homogeneity of independent estimates of the population correction coefficient.

Conditions for the application of χ^2-test:

1. The sample observations should be independent

2.Constraints on the cell frequencies if any must be linear

3. No.of the total frequency should be atleast 50

4. No theoretical cell frequency should be less than 5

Example:14

The number of automobile accidents per week in a certain community are as follows. 12,8,20,2,14,10,15,6,9,4. Are these frequencies in agreement with the belief that accident conditions were the same during this 10 week period.

Solution:

Expected frequency of accidents each week =100/10 = 10

Null hypothesis Ho: The accident conditions were the same during the 10 week period.

Observed Frequency	Expected Frequency	(0- E)	(0- E)^2 /E
12	10	2	0.4
8	10	-2	0.4
10	10	10	10.0
2	10	-8	6.4
14	10	4	1.6
10	10	0	0
15	10	5	3.5
6	10	-4	1.6
9	10	-1	0.1
4	10	-6	3.6

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] = 26.6$$

Tabulated value of χ^2 at 9 degrees of freedom is 16.9.

Calculated χ^2 >Tabulated χ^2

$\therefore$ H_0 is rejected.

Exercise :

1. The no.of automobile accidents per week in a certain community are as follws: 12, 8, 20, 2, 14, 10, 15, 6, 9, 4. Are these frequencies in agreement with the belief that accident condition, were the same during this 10 week period.

2. The following figures show the distribution of digits in numbers chosen at random from a telephone directory .

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Test whether the digits may be taken to occur equally frequently?

Chi squire test for independence of attributes:

An attribute means a quality or characteristic. Let us consider two attributes A and B. A is divided into two classes and B is divided into two classes. The various cell frequencies can be expressed in the following table known as 2 X 2 contingency table.

A	a	B
B	c	D

a	b	a+b
c	d	c+d
a+c	b+d	N

The expected frequencies are given by,

$E(a)=\frac{(a+c)(a+b)}{N}$	$E(a)=\frac{(b+d)(a+b)}{N}$	a+b
$E(a)=\frac{(a+c)(c+d)}{N}$	$E(a)=\frac{(b+d)(c+d)}{N}$	c+d
a+c	b+d	N

H_0 : Attributes are independent

Degrees of freedom = (r-1)(c-1)

r = no of rows

c = no of columns

Example:15

The following table gives the classification of 100 workers according to the sex and nature of work. Test whether the nature of work is independent of the sex of the worker.

	Stable	Unstable	Total
Males	40	20	60
Females	10	30	40
Total	50	50	100

Solution:

1. The parameter of interest is χ^2

2. H_0 : Nature of work is independent of the sex of the workers.

3. H_1 : Nature of work is not independent of the sex of the workers.

4. $\alpha = 0.05$ d.f = (r-1)(c-1)=1

5. Reject H_0 if $\chi^2 > 3.841$ at 5%

6. Computation:

Expected frequencies are given in the table.

$\frac{50\times 60}{100} = 30$	$\frac{50\times 60}{100} = 30$	60
$\frac{50\times 40}{100} = 20$	$\frac{50\times 40}{100} = 20$	40
50	50	100

Calculation of χ^2:

Observed Frequency	Expected Frequency	(0- E)	(0- E)^2 /E
40	30	100	3.33
20	30	100	3.33
10	20	100	5

30	20	100	5

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] = 16.66$$

7. Conclusion: Tabulated value of χ^2 for 1 degrees of freedom at 5% level of significance is 3.84

Calculated χ^2 >Tabulated χ^2

$\therefore$ H_0 is rejected. We conclude that the nature of the workers are not independent.

Example:16

Out of 8000 graduates in a town 800 are females, out of 1600 graduate employees 120 are females, Use χ^2 to determine if any distinction is made in appointment on the basis of sex. Value of χ^2 at 5% level for one degree of freedom is 3.84.

Solution:

	Female	Male	Total
Graduates	800	7200	8000
Employees	120	1480	1600
Total	920	8680	9600

1. The parameter of interest is χ^2

2. H_0 : No difference between 2 treatments

3. H_1 : Difference between 2 treatments

4. $\alpha = 0.05$ d.f = (r-1)(c-1)=1

5. The test statistic $\chi^2 = \dfrac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(b+d)(c+d)}$

6. Reject H_0 if $\chi^2 > 3.841$ at 5%

7. Computation: $\chi^2 = \dfrac{(800X1480 - 7200X120)^2(9600)}{920X8680X8000X1600} = 9.617$

8. Conclusion: Tabulated value of χ^2 for 1 degrees of freedom at 5% level of significance is 3.84

Calculated χ^2 >Tabulated χ^2

$\therefore H_0$ is rejected. We conclude that the treatment are not independent

Exercise:

On the basis of information given below about the treatment of 200 patients suffering from a disease, state whether the treatment is comparatively superior to the conventional treatment.

	Favourable	Not favourable	Total
New	60	30	90
Conventional	40	70	110

S.NO	Questions	OPT 1	OPT 2
1	The population consisting of all real numbers in an example of ____	An infinite population	An finite population
2	The propability distribution of a statistic is called _____	normal distribution	Sampling distribution
3	A part selected from the population is called a _	sample	Population mean
4	_____is the standard deviation of the sampl	standard error	Population mean
5	The chi square test was devised by _____	Fisher	gauss
6	Null hypothesis is the hypothesis of _____	difference	mean
7	Alternative hypothesis complementary to_____	hypothesis	testing of hypothesis
8	Type I error is committed when the hypothesis is true but our test___it	rejects	accept
9	A Type II error is made when_____	the null hypothesis is accepted when it is false.	the null hypothesis is rejected when it is true.
10	The best critical region consists of _____	extreme positive values	extreme negative values
11	The standard deviation of sampling distribution is called_____	standard error	mean error
12	Standard error provides an idea about the _____of sample	unreliability	normality

OPT3	**OPT 4**	**OPT 5**	**OPT 6**	**ANSWERS**
sample	normal			An infinite population
binomial distribution	Sample			Sampling distribution
error	mean square			sample
sample	sampling			standard error
laplace	karl pearson			karl person
no difference	variance			no difference
null hypothesis	Type-I			null hypothesis
null hypthesis	alternative hypothesis			rejects
the alternate hypothesis is accepted when it is false.	the null hypothesis is accepted when it is true.			the null hypothesis is accepted when it is false.
both (a) and (b)	neither (a) nor (b)			both (a) and (b)
error	variance			standard error
reliability	simple			unreliability

13	Normal distribution is a limiting form of_____distribution	Binomial	normal
14	A hypothesis may be classified as_____	Simple	Composite
15	The standard normal distribution is also known as _____distribution	unit normal	normal
16	if v tends to infinity , the chi-square distrbution tends to _____ distribution	normal distribution	Sampling distribution
17	The mean of sampling distribution of means is equal to the_____	Mean	Population mean
18	If a test of hypothesis has a Type I error probability (α) of 0.01, we mean____	if the null hypothesis is true, we don\'t reject it 1% of the time.	if the null hypothesis is true, we reject it 1% of the time.
19	Students t- test is applicable only when____	the variate values are independent	the variate is distributed normally
20	A contingencies table should have frequencies in ____	percentages	proporation
21	Student's t- test is applicable in case of_____	Small samples	for samples of size between 5 and 30
22	Which distribution is used to test the equality of population means ____	chi-square distribution	F-Distribution
23	The shape of t-distribution is similar to that of _____	chi-square distribution	F-Distribution
24	The number of degrees of freedom for contingency table are on the basis of _____	8	4

uniform	sample			Binomial
null	all the above			all the above
uniform	sample			unit normal
binomial distribution	Sample			normal distribution
variance	Sample mean			Population mean
if the null hypothesis is false, we don\'t reject it 1% of the time.	if the null hypothesis is false, we reject it 1% of the time.			if the null hypothesis is true, we reject it 1% of the time.
the sample is not large	all the above			all the above
frequencies	ratio			frequencies
large samples	all the above			Small samples
Normal distribution	t- distribution			F-Distribution
Normal distribution	uniform distribution			chi-square distribution
3	2			4

25	Student's t- test was invented by_____	R.A.Fisher	G.W.Snedector
26	The degrees of freedom for contingency table are on the basis of _____	n-1	r-1
27	The calculated value of chi-square is : _____	always positive	always negative
28	Degrees of freedom for statistic chi-square in case of contingency table of order (2x2) is_____	3	4
29	Normal distribution is applicable in case of_____	Small samples	for samples of size between 5 and 30
30	Degrees of freedom for chi-square in case of contingency table of order (4x3) are _____	12	9

W.S.Gosset	W.G.Cochran			W.S.Gosset
c-1	r-2			r-1
either positive or negative	none of these			always positive
2	1			1
large samples	all the above			large samples
8	6			6

	The term "Analysis of variance was introduced by" _____	R.A. FISHR	anova	Roma	aggarwal			R.A. FISHR
	The assumptions in analysis of variance are the same as____	F-test	T-Test	Anova	mean square			mean square
	Anova table stands for _____ ___	Variance Table	Analysis of variance Table	Analysis	Radom variable			Analysis of variance Table
	The Science of experimental designs is associated with the name_ ____	Latin square	Random block design	Latin cubes	absolute			Latin square

	If degress of freedom increase , _____ decreases	quardrar	skewness	curve	normal			skewness
	All the odd moments about the mean are _____	zero	one	two	two are more			zero
	Analysis of variance utilises _____	F-test	chi-Square test	Z-test	t-test			F-test
	The mean of the chi-square distribution is equal to the _____ freedom	mode	standard deviation	degrees	sample			degrees
	The degree of freedom for t distrbution _____	size of sample -1	sample	Normal distribution	curve			size of sample -1

UNIT – IV

Design of Experiments

Introduction:

Experimental are a nature part of the engineering and scientific decision making process. Designed experiments play a very important role in engineering design and development and in the improvement of manufacturing processes.

Definition :

The logical construction of the experiment in which the degree of uncertainty with which the degree of uncertainty with which the inference is drawn may be well defined.

Basic Principles of Experimental Design :

According to Prof. Ronald A Fisher, the basic principles of the design of experiments are :

1. Replication
2. Randomisation
3. Local control

1. Replication:

The repetition of the treatment under investigation, which results in more reliable estimate than is possible with a single observation.

2. Randomisation:

Randomisation is a process of assigning the treatment to various experimental units in a purely chance manner. It insures that different treatment, by the repetition of the experiment, on the average are subject to equal environmental effect. Randomisation eliminates bias in any form.

3. Local control :

The process of reducing the experimental error by dividing the relatively heterogeneous experimental area into homogeneous blocks is known as local control.

Analysis of variance :

The analysis of variance is a powerful statistical tool for tests of significance. The test of significance based on t distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation when we have three or more samples to consider a time an alternative procedure is needed. The answer to this problem is provided by

the technique of analysis of variance. Thus basic purpose of analysis of variance is to test the homogeneity of several means.

The term 'Analysis of variance' was introduced by Prof .R.A.Fisher. Variations 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 Assignable causes and chance causes. The variation due to assignable causes can be detected and measured whereas the variation due to chance cause is beyond the control of human hand and cannot be traced separately.

Definition : (**ANOVA)**

Separation of variance ascribable to one group of causes from the variance ascribable to other group.

Assumptions :

For the validity of F-Test in ANOVA the following assumptions are made :

(i) The observations are independent

(ii) Parent population from which observations are taken is normal, and

(iii) Various treatment and environmental effects are additive in nature.

SSC – between sum of squares (column)

TSS - total sum of squares

SST - sum of squares due to treatments

MSS – mean sum of squares

SSE – error sum of squares (or) with in sum of squares

RSS – row sum of squares

C.F – correlation factor

C.D – critical difference

SSR – sum of squares between rows

MSC – mean sum of squares (with in columns)

MSR - mean sum of squares (between rows)

N - number of observations

N_1 – no. of elements in each column

N_2 - no. of elements in each row

One way classification:

One way classification observations are classified according to one factor. This is

exhibited column wise.

ANOVA Table for one-way classified data:

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	Variance ratio
Between columns Within columns (Error)	C-1 N-C	SSC SSE	$MSC = \frac{SSC}{C-1}$ $MSE = \frac{SSE}{N-1}$	$F = \frac{MSC}{MSE}$ (or) $F = \frac{MSE}{MSC}$
Total	N-1	TSS		

The F ratio should be calculated in such way that F>1.

Completely Randomised Design (CRD)

The CRD is the simplest of all the designs, based on the principles of randomisation and replication. In this design the treatment are allocated at random to the experimental units over the entire experimental material.

Advantages:

1. CRD results in the maximum use of the experimental units since all the experimental material can be used.
2. The design is very flexible.
3. The statistical analysis remains simple if some or all the observations for any treatment are rejected or lost or missing for some purely random accidental reason.
4. It provides the maximum number of degrees of freedom for the estimation of the error variance, which increases the sensitivity or the precision of the experiment for small treatments.

Disadvantages :

In certain circumstances the design suffers from the disadvantage of being inherently less informative than other more sophisticated layouts. This usually happens if the experimental material is not homogeneous.

Applications:

1. CRD is most useful in laboratory technique and methodological studies is physics, chemistry or cookery, in chemical and biological experiments.

Example: In Physics , Chemistry or cookery in chemical and biological experiments ,

in some green house studies etc, where either the experimental material is homogeneous.

2. CRD is recommended in situations where an appreciable fraction of units is likely to be destroyed or to fail to respond.

Note:

The mathematical model and statistical analysis for CRD is same as that of one way classification.

Working Procedure:

1. H_0 : There is no significant difference between the treatments.
2. H_1 : There is significant difference between the treatments.

Steps:

1. Find N, the no.of observations
2. Find T, the total value of all observations
3. Find $\frac{T^2}{N}$, the correction factor.
4. Calculate the total sum of squares TSS= $\sum X_1^{\,2} + \ldots\ldots - \frac{T^2}{N}$
5. Calculate the column sum of squares SSC = $\left(\frac{\sum X_1}{N_1}\right)^2 + \ldots\ldots - \frac{T^2}{N}$

   Here N_1 is no.of elements in each column SSE= TSS – SSC
6. Prepare the ANOVA table to calculate F-ratio.
7. Find the table value.
8. Conclusion.

Example : 1

A test was given to five students taken at random from the fifth class of three schools of a town. The individual scores are

School I : 9 7 6 5 8

School II : 7 4 5 4 5

School III : 6 5 6 7 6

Carry out the analysis of variance and state your conclusion.

Solution:

H_o : There is no significant difference between the 3 schools

H_1 : There is a significant difference between the 3 schools

X_1	X_2	X_3	Total			
9	7	6	22	81	49	36
7	4	5	16	49	16	25
6	5	6	17	36	25	36
5	4	7	16	25	16	49
8	5	6	19	64	25	36
35	25	30	90	255	131	182

Step:

1. N = 15

2. T = 90

3. C.F = $\frac{T^2}{N}$ = 540

4. TSS = $\sum X_1^2 + \sum X_2^2 + \sum X_3^2 - \frac{T^2}{N}$

= 255+131+182-540

=28

5. SSC = $\frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_1} + \frac{(\sum X_3)^2}{N_1} - \frac{T^2}{N}$

= 245+125+180-540

=10

SSE = TSS-SSC

= 28-10 = 18

6. ANOVA Table

Source of variation	Degrees of freedom	Sum of squares	Mean SS	Variance ratio (F)	Table value 5% level
b/w column	c-1 = 2	SSC = 10	MSC= 5	$F_C = \frac{MSC}{MSE} = 3.33$	$F_C(2,12)=3.89$
Error	N-c = 12	SSE = 18	MSE = 1.5		

7. Conclusion:

Cal F_C < Tab F_C, So we accept H_0 at 5% level of significance.

Example :2

The following table shows some of the results of an experiment on the effect of applications of sulphur in reducing scale disease of potatoes. The object in applying sulphur is to increase the acidity of the soil since scale does not thrive in very acid soil. Both a spring and a fall application of each treatment was tested so that in all there were 7 distinct treatments. The quality to be analysed in the scale index.

F_3	0	S_6	F_{12}	S_6	S_{12}	S_3	F_6
9	12	18	10	24	17	30	16
O	S_3	F_{12}	F_6	S_3	0	0	S_6
10	7	4	10	21	24	29	12
F_3	S_{12}	F_6	0	F_6	S_{12}	F_3	F_{12}
9	7	18	30	18	16	16	4
S_3	0	S_{12}	S_6	0	F_{12}	0	F_3
9	18	17	19	32	5	26	4

F = Full, S = Spring application, O = Control. The numbers 3,6,12 are the amount of sulphur in 100 lb per acre.

Analyse the experiment and give in detail your conclusion

Solution:

H_0 : There is no significant difference between the three scale index.

H_1 : There is significant difference between the three scale index.

	0		F3	S3	F6	S6	F12	S12
	12	30	9	30	16	18	10	17
	10	18	9	7	10	24	4	7
	24	32	16	21	18	12	4	16
	29	26	4	9	18	19	5	17
Total		181	38	67	62	73	23	57
Means		22.6	9.5	16.8	15.5	18.2	5.8	14.2

Step:

1. N = 32

2. T = 181+38+67+62+73+23+57= 501

3. C.F = $\frac{T^2}{N}$ = 7843.78

4. TSS = $\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + ...\sum X_7^2 - \frac{T^2}{N}$

= 9939-7848.78

=2096.22

5. SSC = $\frac{(\sum X_1)^2}{N_1} + ... + \frac{(\sum X_7)^2}{N_1} - \frac{T^2}{N}$

=972.3

SSE = TSS-SSC

= 2096.22 – 972.3 = 1123.92

ANOVA Table

Source of variation	Degree of freedom	Sum of Squares	Mean sum of Squares	Variance Ratio F	Table value at 5% level
b/w column	c-1 = 6	SSC= 972.3	MSC = 162.0	F_C=3.61	$F_C(6,25)$ = 2.49
Error	N-c = 25	SSC = 1123.9	MSE = 44.9		
Total	31	2095.2			

$F_{0.05}(6,25)$ = 2.49 since table value is less than the calculated value.

7. Conclusion:

So we reject H0 at 5% level of significance. All treatments are not same.

Example:3

There are three main brands of a certain powder. A set of 120 sample values is examined and found to be allocated among four groups (A, B, C, D) and three brands (I,II,III) as shown here under:

Brands	Groups			
	A	B	C	D
I	0	4	8	15
II	5	8	13	6
III	8	19	11	13

Is there any significant difference in brands preference in brands :Answer at 55 level.

Solution:

H_o : There is no significant difference between the groups and brands.

H_1 : There is significant difference between the groups and brands.

Brands	Groups

	X_1	X_2	X_3	X_4	Total	X_1^2	X_2^2	X_3^2	X_4^2
Y_1	0	4	8	15	27	0	16	64	225
Y_2	5	8	13	6	32	25	64	169	36
Y_3	8	19	11	13	51	64	361	121	169
Treatment Total Ti	13	31	32	34	110	89	441	354	430

Steps:

1. N=12, N_1=4

2. T=110

3. CF = $\frac{T^2}{N} = \frac{(110)^2}{12}$ = 1008.3

4. TSS = $\sum X_1^{\,2} + \ldots\ldots \sum X_4^{\,2} - \frac{T^2}{N}$

= 89+441+354+430-1008.3

=305.7

5. $SSR = \left(\frac{\sum Y_1}{N_1}\right)^2 + \left(\frac{\sum Y_2}{N_1}\right)^2 + \ldots \left(\frac{\sum Y_3}{N_1}\right)^2 - \frac{T^2}{N}$

= $\frac{(27)^2}{4} + \frac{(32)^2}{4} + \frac{(51)^2}{4} - 1008.3$

= 80.2

6. SSE = TSS - SSR

= 305.7 – 80.2

= 225.50

ANOVA Table:

Source of variation	Degree of freedom	Sum of squares	Mean sum squares	Variance Ratio F	Table value at 5% level
b/w rows (Blocks)	2	SSR = 80.2	MSR = 40.1	F_R = 1.999	F_R (2, 6) = 4.26
Error	12-3 = 9	SSE = 225.5	MSE = 20.06		

7. Conclusion:

Cal F_R < Tab F_R , so H_0 is accepted .

Exercise:

1. The following data shows the lives in hours of four batches of electric lamps.

Batches

A. 1600 1610 1650 1680 1700 1720 1800

B. 1580 1640 1640 1700 1750

C. 1450 1550 1600 1620 1640 1660 1740 1820

D. 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.

2. Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of output are made :

A : 10 12 13 11 10 14 15 13

B : 9 11 10 12 13

C : 11 10 15 14 12 13

Carry out the analysis of variance and state your conclusion.

3.The following data gives the yields on 12 sample plots under three varieties of seed.

A	B	C
21	20	28
23	17	22
16	15	28
20	23	32

Find out if the average yields of land under different variance show significant differences.

Two way Classification :

In two way classification of analysis of variance. We consider one classification along column wise and the other along row wise.

Randomised Block Design: (RBD)

If the treatment are applied at random to relatively homogeneous units within each strata or block and replicated over all the blocks, the design is Randomised block design (RBD).

Advantages :

1. Accuracy
2. Flexibility
3. Ease of Analysis

Disadvantages:

RBD is not suitable for large number of treatments or for cases in which complete block contains considerable variability.

ANOVA Table for two- way classification data:

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	Variance ratio
Column Treatments	c-1	SSC	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$
Row treatment (Blocks)	r-1	SSR	$MSR = \frac{SSR}{r-1}$	$F_R = \frac{MSR}{MSE}$
Error	(c-1) x (r-1)	SSE	$MSE = \frac{SSE}{(c-1)(r-1)}$	
Total	rc-1	TSS		

Thus if an observed value of F obtained is greater then the tabulated value of F for (c-1),(c-1)(r-1) and (r-1), (r-1)(c-1) degrees of freedom at specified level of significance (usually 5% or 1%) then Ho is rejected at that level.

Working rule for two way classification:

1. H_0 : There is no significant difference between the treatments.
2. H_1 : There is significant difference between the treatments.

Steps:

1. Find N, the no.of observations
2. Find T, the total value of all observations
3. Find $\frac{T^2}{N}$, the correction factor.

4. Calculate the total sum of squares TSS= $\sum X_1^{\,2} + \ldots\ldots - \frac{T^2}{N}$

5. Calculate the column sum of squares SSC = $\left(\frac{\sum X_1}{N_1}\right)^2 + \ldots\ldots - \frac{T^2}{N}$

6. Calculate the row sum of squares SSR = $\left(\frac{\sum Y_1}{N_1}\right)^2 + \ldots\ldots - \frac{T^2}{N}$

7. SSE = TSS- SSC -SSR

8. Prepare the ANOVA table to calculate F_C & F_R

9. Find the table value.

10. Conclusion.

Example :4

Consider the results given in the following table for an experiment involving six treatment in four randomised blocks. The treatments are indicated by numbers within parenthesis

Blocks	Treatment and yield					
1	(1)	(3)	(2)	(4)	(5)	(6)
	24.7	27.7	20.6	16.2	16.2	24.9
2	(3)	(2)	(1)	(4)	(6)	(5)
	22.7	28.8	27.3	15	22.5	17.0
3	(6)	(4)	(1)	(3)	(2)	(5)
	26.3	19.6	38.5	36.8	39.5	15.4
4	(5)	(2)	(1)	(4)	(3)	(6)
	17.7	31.0	28.5	14.1	34.9	22.6

Test whether the treatments differ significantly.

Solution:

H_o : There is no significant difference between the treatments and blocks.

H_1 : There is significant difference between the treatments and blocks

Blocks	Treatments						Block Total
	(1)	(2)	(3)	(4)	(5)	(6)	Bj
1	24.7	20.6	27.7	16.2	16.2	24.9	130.0

2	27.3	28.8	22.7	15.0	17.0	22.5	133.3
3	38.5	39.5	36.8	19.6	15.4	26.3	176.1
4	28.5	31.0	34.9	14.1	17.7	22.6	148.8
Treatment Total Ti	119.0	119.9	123.3	64.9	66.3	96.3	588.7

Steps:

1. N=24

2. T=588.7

3. CF = $\frac{T^2}{N} = \frac{(588.7)^2}{24}$ = 14440.32

4. TSS = $\sum X_1^{\,2} + \ldots\ldots \sum X_6^{\,2} - \frac{T^2}{N}$

= 1349.57

5. $SSC = \left(\frac{\sum X_1}{N_1}\right)^2 + \ldots\ldots \left(\frac{\sum X_6}{N_1}\right)^2 - \frac{T^2}{N}$

= 903.59

6. $SSR = \left(\frac{\sum Y_1}{N_2}\right)^2 + \left(\frac{\sum Y_2}{N_2}\right)^2 + \ldots \left(\frac{\sum Y_4}{N_2}\right)^2 - \frac{T^2}{N}$

= 218.6

7. SSE = TSS-SSC-SSR

= 227.47

ANOVA Table:

Source of variation	Degree of freedom	Sum of squares	Mean sum squares	Variance Ratio F	Table vaiue at 5% level
b/w column(Treatment)	5	SSC= 903.59	MSC = 180.72	F_C = 11.92	F_C (5, 15) = 4.5
b/w rows (Blocks)	3	SSR = 218.6	MSR = 72.86	F_R = 4.8	
Error	15	227.47	MSE = 15.16		F_R (3, 15) = 5.42

9. Conclusion:

Cal F_C > Tab F_C , so H_0 is rejected.

Cal F_R < Tab F_R , so H_0 is accepted. Hence we conclude that treatments effects are not alike where as the blocks & parenthesis.

Example :5

The yield of four strains of a particular variety of wheat was planted in five randomized blocks in kgs per plots is given below. Test for difference between blocks and difference between strains.

		Blocks				
		1	2	3	4	5
Strains	A	32	34	34	35	36
	B	33	33	36	37	34
	C	30	35	35	32	35
	D	39	22	30	28	28

Solution:

H_o : There is no significant difference between the strains and blocks.

H_1 : There is significant difference between the strains and blocks

Subtract 30 from each value we get

Blocks				Treatments						
	X1	X2	X3	X4	X5	Total	X_1^2	X_2^2	X_3^2	X_4^2	X_5^2
Y_1	2	4	4	5	6	21	4	16	16	25	36
Y_2	3	3	6	7	4	23	9	9	36	49	16
Y_3	0	5	5	2	5	17	0	25	25	4	25
Y_4	-1	-8	0	-2	-2	-13	1	64	0	4	4
Treatment Total Ti	4	4	15	12	13	48	14	114	77	82	81

Steps:

1. N=20, N_1=4

2. T=48

3. CF = $\frac{T^2}{N} = \frac{(588.7)^2}{24} = 115.2$

4. TSS = $\sum X_1^{\,2} + \ldots\ldots \sum X_6^{\,2} - \frac{T^2}{N}$

= 14+114+77+82+81-115.2

=252.8

$$5.SSC = \left(\frac{\sum X_1}{N_1}\right)^2 + ...\left(\frac{\sum X_6}{N_1}\right)^2 - \frac{T^2}{N}$$

= 27.3

6. $$SSR = \left(\frac{\sum Y_1}{N_2}\right)^2 + \left(\frac{\sum Y_2}{N_2}\right)^2 + ...\left(\frac{\sum Y_4}{N_2}\right)^2 - \frac{T^2}{N}$$

= 170.4

7. SSE = TSS-SSC-SSR

= 252.8-27.3-170.4

= 55.1

ANOVA Table:

Source of variation	Degree of freedom	Sum of squares	Mean sum squares	Variance Ratio F	Table value at 5% level
b/w column(Treatment)	4	SSC= 27.3	MSC = 6.825	F_C = 1.49	F_C (4, 12) = 3.2
b/w rows (Blocks)	3	SSR = 170.4	MSR = 56.8	F_R = 12.37	
Error	12	SSE=55.1	MSE = 4.59		F_R (3, 12) = 3.4

9. Conclusion:

Cal F_C < Tab F_C , so H_0 is accepted .

Cal F_R > Tab F_R , so H_0 is rejected.

Exercise:

1. A coffee company appoints four salesmen P,Q,R and S and observes their sales in three districts A,B and C. The figures given below(Figures are in lakhs of rupees.

	Salesmen				
District	P	Q	R	S	District Total
A	36	36	21	35	128
B	28	29	31	32	120

C	26	28	29	29	112
Salesmen Totals	90	93	81	36	

Carryout analysis of variance.

2. A following data represent the no. of. units production per day turned out by different workers using four different types of machines.

	Machine type			
Workers	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

Carry out ANOVA.

Latin Square Design : (LSD)

In RBD whole of the experimental area is divided into relatively homogeneous groups and treatments are allocated at random to units within each block i.e., randomisation was restricted within blocks. A useful method to eliminate fertility variations consists in an experimental layout which will control variation in two perpendicular directions, such a layout is a lain square design.

In this design, the number of treatment is equal to the number of replications. Thus in case of m treatments, there have to m x m = m^2 experimental units.

Latin square:

The data are classified according to columns , rows and varieties and are arranged in a square known as latin square.

Advantages:

1.With two way grouping or stratification LSD controls more of the

variation than CR or RBD.

2. The statistical analysis is simple.

3. More than one factor can be investigated simultaneously and with fewer trials than more complicated designs.

4. LSD controls variability in two directions of the experimental material.

5. The analysis of the design in simple and straight forward and is a three way classification of ANOVA.

Disadvantages :

1. The fundamental assumption that there is no interaction between different factors May not be true in general.
2. Unlike RBD in LSD the number of treatments is restricted to the number of replications and this limits its field of application.
3. In the field layout RBD is much easy to manage than LSD.

4.The process of randomization is not as simple as in RBD.

5. The no.of treatments should be equal to the no.of rows and no.of columns.

6. The experimental area should be in the form of a square.

7. It is suitable only in the case of smaller no.of treatments.

8. A 2x2 latin square is not possible.

WORKING RULE:

The analysis is done in a way similar to two way classification. The different sum of squares are obtained as follows:

Steps:

1. Find N
2. Find T
3. Find $\frac{T^2}{N}$
4. Find TSS
5. Find SSC
6. Find SSR and SSK

7.ANOVA table

8.Conclusion.

ANOVA Table:

Source of	Degrees of	Sum of squares	Mean sum	F

variation	freedom		squares	
Between Rows	k-1	SSR	$MSR = \frac{SSR}{k-1}$	$F_R = \frac{MSR}{MSE}$
Between Columns	k-1	SSC	$MSC = \frac{SSC}{k-1}$	$F_C = \frac{MSC}{MSE}$
Between Treatments	k-1	SSK	$MSK = \frac{SSK}{k-1}$	$F_T = \frac{MSK}{MSE}$
Error	(k-1) (k-2)	SSE	$MSE = \frac{SSE}{(k-2)(k-1)}$	
Total	k^2-1			

The variance ratios F_R, F_C, F_T are calculated in a such way that they are each greater than one.

Note:

$\frac{MSC}{MSE} < 1$ then take $F_C = \frac{MSE}{MSC}$

Example : 6

The following is a Latin square of a design when 4 varieties of seeds are being tested. Set up the analysis of variance table and state your conclusion. You may carry out suitable change of origin and scale.

A	105	B	95	C	125	D	115
C	115	D	125	A	105	B	105
D	115	C	95	B	105	A	115
B	95	A	135	D	95	C	115

Solution:

Subtract 100 and then divided by 5 we get

A	1	B	-1	C	5	D	3
C	3	D	5	A	1	B	1
D	3	C	-1	B	1	A	3
B	-1	A	7	D	-1	C	3

Brand	Four varieties of seed								
	X_1	X_2	X_3	X_4	Total	X_1^2	X_2^2	X_3^2	X_4^2
Y_1	1	-1	5	3	8	1	1	25	9
Y_2	3	5	1	1	10	9	25	1	1
Y_3	3	-1	1	3	6	9	1	1	9
Y_4	-1	7	-1	3	8	1	49	1	9

Treatment Total Ti	6	10	6	10	32	20	76	28	28

Steps:

1. N=16, N_1=4

2. T=32

3. CF = $\frac{T^2}{N} = \frac{(32)^2}{16} = 64$

4. TSS = $\sum X_1^{\,2} + \ldots\ldots \sum X_4^{\,2} - \frac{T^2}{N}$

= 20+76+28+28-64

=88

5. $SSC = \left(\frac{\sum X_1}{N_1}\right)^2 + \ldots\ldots + \left(\frac{\sum X_4}{N_1}\right)^2 - \frac{T^2}{N}$

= 9+25+9+25-64

= 4

6. $SSR = \left(\frac{\sum Y_1}{N_2}\right)^2 + \ldots\ldots + \left(\frac{\sum Y_4}{N_2}\right)^2 - \frac{T^2}{N}$

$= \frac{(8)^2}{4} + \frac{(10)^2}{4} + \frac{(6)^2}{4} + \frac{(8)^2}{4} - 64$

= 16+25+9+16-64

= 2

7. To find SSK:

Arrange the elements in the order of treatment.

A	1	1	3	7	12
B	-1	1	1	-1	0
C	5	3	-1	3	10
D	3	5	3	-1	10

$SSK = \frac{(12)^2}{4} + \frac{(10)^2}{4} + \frac{(10)^2}{4} - 64$

= 36+0+25+25 -64

= 22

8. SSE = TSS – SSC-SSR-SSK

= 88-4-2-22

= 60

ANOVA Table:

Source of variation	Degree of freedom	Sum of squares	Mean sum squares	Variance Ratio F	Table value at 5% level
b/w rows (Blocks)	k-1=3	SSR = 2	MSR = 0.67	F_R = 14.9	F_R (6, 3) = 8.94
b/w Columns	k-1=3	SSC = 4	MSC = 1.33	F_C=7.52	F_C (6, 3) =8.94
b/w treatments	k-1=3	SSK = 22	MSK = 7.33	F_T=1.36	F_T (6, 3) =8.94
Error	(k-1)(k-2) =6	SSE = 60	MSE = 10		

9. Conclusion:

Cal F > Tab F , There is a significant difference between rows as well as between columns. But there is no significant difference between treatments. so H_0 is accepted .

Exercise :

1. Set up the analysis of variance for the following results of a Latin Square design.

A	**B**	**C**	**D**
12	**19**	**10**	**8**
C	**B**	**D**	**A**
18	**12**	**6**	**7**
B	**D**	**A**	**C**
22	**10**	**5**	**21**
D	**A**	**C**	**B**
12	**7**	**27**	**17**

2. Analyze the following results of a Latin square experiments.

	1	**2**	**3**	**4**
1	**A (12)**	**D (20)**	**C (16)**	**B (10)**
2	**D (18)**	**A (14)**	**B (11)**	**C (14)**

3	**B (12)**	**C (15)**	**D (19)**	**A (13)**
4	**C (16)**	**B (11)**	**A (15)**	**D (20)**

The letters A,B,C,D denote the treatment and the figure is brackets denote the observation.

S.NO	Questions	OPT 1	OPT 2	OPT3
1	The most widely used of all experimental design is _____	Randomised block design	latin square	Mean square of error
2	The experimental area should be in the form of _____	Circle	parabola	square
3	The word _____ in analysis of variance is used to refer to any factor in the experiment.	Error	normality	treatment
4	In the case of one-way classification the total variation can be split into _____	Two components	Three components	Four components
5	Analysis of variance can be used when there are samples of _____ sizes	Equal	unequal	greater than
6	Mean square of error =_____ for one way classification	SSE	SSE\n-c	SSE\1-c
7	Total variation SST = _____ for one way classification	SSC+SSR	SSE+TSS	SSC+SSE
8	Mean square between column mean _____	SSE	SSE\n-c	SSE\c-1
9	_____ stands for mean square between samples	MSC	SSE	SSR
10	The stimulus to the development of theory and practice of experimental design came from_____	Agrarian research	industry research	astronomy research

OPT 4	**OPT 5**	**OPT 6**	**ANSWERS**
experimental error			Randomised block design
ellipse			square
Homogencity			treatment
Only one component			Two components
less than			unequal
SSE\r-1			SSE\n-c
TSS			SSC+SSE
SSE\r-1			SSE\c-1
SST			MSC
medicine research			Agrarian research

11	The analysis of variance originated in _____	Agrarian research	industry research	astronomy research
12	The Latin square model assumes that interactions between treatment and row and column groupings are _____	Existent	non-existent	experimental error
13	The science of experimental designs is associated with the name _____	Randomised block design	latin square	Mean square of error
14	In 4×4 Latin square, the total of such possibilities are_____	8	10	200
15	The latin sqares are most widely used in the field of _____	Agriculture	industry	astronomy
16	The total number of possibilities in which arrangements can be made in 3×3 Latin square are_____	6	9	12
17	The one way classification is exhibited _____ wise	Row	column	both a & b
18	The shape of the experimental material should be _____	Circle	parabola	rectangular
19	The number of treatments should be _____ number of rows and number of columns	Equal	unequal	greater than
20	Latin square design controls variability in _____ directions of the experimental material	One	two	three
21	_____ Latin square is not possible	2×2	3×3	4×4
22	In the case of two-way classification, the total variation (TSS) equals_____	SSR + SSC + SSE	SSR -SSC + SSE	SSR + SSC – SSE

medicine research			Agrarian research
Mean square of error			non-existent
None of these			latin square
576			576
medicine			Agriculture
120			12
None of these			both a & b
ellipse			rectangular
less than			Equal
four			two
5×5			2×2
SSR + SSC			SSR + SSC + SSE

23	The assumptions in analysis of variance _____	Normality	Homogencity	independence of error
24	In one way classification the dates are classified according to _____ factor	One	two	three
25	Equality of several normal population means can be tested by _____	Bartlet' s test	F - test	chi square-test
26	Analysis of variance technique was developed by_____	S. D. Poisson	Karl – Pearson	R.A. Fisher
27	Analysis of variance technique originated in the field of_____	Agriculture	industry	astronomy
28	One of the assumption of analysis of variance is that the population from which the samples are drawn is_____	Binomial	Poisson	Chi-square
29	In a two way classification the datas are classified to _____ factor	One	two	three
30	In the case of one-way classification with N observations and t treatments, the error degrees of freedom is_____	N-1	t -1	N-t
31	In the case of one-way classification with t treatments, the mean sum of squares for treatment is_____	SST/N-1	SST/ t-1	SST/N-t
32	In the case of two-way classification with r rows and c columns, the degrees of freedom for error is_____	(rc) – 1	(r-1).c	(r-1) (c-1)
33	Latin square design controls variability in _____ directions of the experimental material	One	two	three

both a,b& c			both a,b& c
four			One
t- test			F - test
W. S. Gosset			R.A. Fisher
medicine			Agriculture
Normal			Normal
four			two
Nt			N- t
SST/t			SST/ t-1
(c-1).r			(r-1) (c-1)
four			two

34	With 90, 35, 25 as TSS, SSR and SSC respectively in case of two way classification, SSE is____	50	40	30
35	One of the assumptions of Analysis of variance is observations are _____	independent	dependent	Industry
36	Total variation in two – way classification can be split into _____ components.	two	three	four
37	In the case of one way classification with 30 observations and 5 treatment, the degrees freedom for SSE is _____	20	19	24
38	In the case of two-way classification with 120, 54, 45 respectively as TSS, SSC, SSE, the SSR is _____	19	21	20
39	The origin of statistics can be traced to _____	State	Commerce	Economics
40	'Statistics may be called the science of counting' is the definition given by _____	Croxton	A.L.Bowley	Boddington
41	_____ is one of the statistical tool plays prominent role in agricultural experiments.	Analysis of variance	Normality	Homogenecity
42	The Latin square model assumes that interactions between treatment and row and column groupings are _____	Existent	non-existent	experimental error
43	In 4×4 Latin square, the total of such possibilities are_____	8	10	200
44	The sum of the squares between samples are denoted by	SSR	SSE	TSS

20			30
Genetics			independent
five			three
25			25
27			21
Industry			State
Webster			A.L.Bowley
independence of error			Analysis of variance
Mean square of error			non-existent
576			576
SSC			SSC

UNIT-V

RELIABILITY AND QUALITY CONTROL

INTRODUCTION:

In every life situation we use the word, "reliability" or "reliable" in the sense of "dependable" or "dependability". Reliability theory is concerned with determining the probability that a system, possibly consisting of many components will function.

CONCEPTS OF RELIABILITY:

Terms related to reliability

1. Reliability 2.Failure 3.Maintainability 4. Availability

1.Reliability may be defined as the probability that a component will perform properly for a specified period of time t under a given set of operating conditions.

If a component is put into operation at some specified time, say t=0 and if T is the time until it fails or ceases to function properly, T is called the life length or time to failure of the component, $T \geq 0$ is obviously a continuous random variable with some pdf of $(f(t))$. Then the reliability of the component at time denoted by R(t) and is defined as

$$R(t) = P(T > t)$$
$$= 1 - P(T \leq t)$$
$$= 1 - F(t)$$

Where F (t) is the cumulative distribution function of T, given by F (t) =

$$R(t) + F(t) = 1$$

Thus R (t) = 1-F (t) =

$$F'(t) = f(t)$$

The component is assumed to be working properly at time t=0. Ie) R(0) =1 and no component can work for ever without failure.

R(t) =0. For t<0, reliability has no meaning but we let R(t) = 1 for t<0. F(t) is called unreliability.

2. Failure: Terms related to failure

1. Failure
2. Failure rate (or) hazard rate
3. Mean time between failures (MTBF)
4. Mean time to failure(MTTF)

i) Failure:

A failure is the partial and total loss or change in those properties of a device (or system) in such a way that its functioning is seriously impeded or completely stopped some components have well defined failures others do not.

ii) Failure rate (or) Hazard rate:

The conditional probability of failure per unit time is given by and is called the instantaneous failure rate or hazard function of the component, denoted by λ(t).

λ(t) =

iii) Mean time between failures (MTBF)

MTBF is normally taken to be the mean time between the and failure in a system when "n" is relatively large MTBF is be a function of time when a system is operated, MTBF will fluctuate and then stabilize.

iv) Mean time failure (MTTF):

MTTF is only for non repairable items. The expected value of the time of failure T, denoted by E(T) and variance of T, denoted by σ_T^2 are 2 important parameters frequently used to characterize reliability. E(T) is called mean time failure and denoted by MTTE.

MTTF=E(T)= =

Var(T) = T = E()-

= f(t)dt –

3. Maintainability:

It can be defined as the probability that failed equipment is restored to operable condition in a specified time (called down time) when the maintenance is performed under state conditions means time to repair

MTTR =

MTTR is the statistical mean time for active repairs.

4. Availability:

Availability is another measure of performance of the maintained equipment

Availability =

Properties:

The reliability R(t) has the properties

1. 0 R(t)1
2. R(0)=1 and R()=0 and
3. R(t) is function of t

`

Note:- 1) MTTF=

Control charts for measurement [x and R charts] - Control charts for attributes [P,C and np chart]

Control chart:

A control chart provides a basis for deciding whether the variation in the output is due to random causes or due to assignable causes

A control chart is due to display successive measurement of a process with a center line and control limits

The control limits are above and below the center line and are equidistant from the center line and are known as upper control limit (UCL) and lower control limit (LCL)

The control chart helps us decide whether the process of production is control is control or not.

TYPES OF CONTROL CHARTS:

1. Control charts for variables
2. Control charts for attributes

Procedure to draw the chart and R chart:

1. The sample value in each of the N sample each of size n will be given. Let ,,……, be the means of the N-samples and ,,…. be the ranges of the samples. By range of a sample we mean the max sample value minus the min sample value in that sample.
2. We compute = (+…..+) and = (+….+)
3. The values of ,, for the given sample size n are taken from the table of control chart constants.
4. Then the values of the control limits and the control limits and are computed
5. On the ordinary graph sheet the sample no's are represented on the x axis and the sample means on the y axis and the sample ranges on the y axis
6. For drawing the mean chart, we draw the 3 lines y =, y = - and y = + which represent the central line, the LCL line & UEL line. Also we plot the points whose co-ordinates are (1,),(2,)….(N,) and join adjacent points by line segments. The graph thus obtained is the - chart.
7. For drawing mean chart, we draw the 3 lines y= ,y= and y= which represent resp, the central line, the LCL and UCL line ,Also we polot the point whose coordinate are (1 (2, (N) and join adjacent point by line segments. The graph thus obtained is the R charts.

Construction of chart:

Draw independent samples each of size n from a large production process .Let by the means of these samples X=.

The control limits are given by UCL= +3.SE()

LCL= -3.SE()

CL=

SE()= being the SD of the production

If is not available the SD of the sample distribution of the mean can be taken as the best estimate of ,In the case of small sample ,the estimate of SE of is . Alternatively in the case of small sample of size less than 20, The control limits are UCL= +

LCL= -

CL =

Range chart (R chart):

For samples of size 20 the Range provides a good estimate of . Here to measure the variance in the variable, Range chart is used,

Construction of R chart:

Let = be the value of ranges in k sample

The control limits are given by LCL=

UCL=

The factors & are determined from statistical table for known sample size

The most common control charts under this category

1. Control chart for no of defectives
2. Control chart for fraction defectives

Problems based on and chart:

Example:1

Given below are the values of samples mean and sample range R for 10 samples, each of size 5, Draw the appropriate mean and range charts and content on the state of control of the process

Sample no	1	2	3	4	5	6	7	8	9	10
Mean	43	49	37	44	45	37	51	46	43	47
Range	5	6	5	7	7	4	8	6	4	6

Solution:

=

= (43+49+…+47)

= 44.2

=

= (5+6+…+6)

=5.8

From the table of control chart for sample size n=5 we have =0.577

= 0

= 2.115

i) Control limits for chart:-

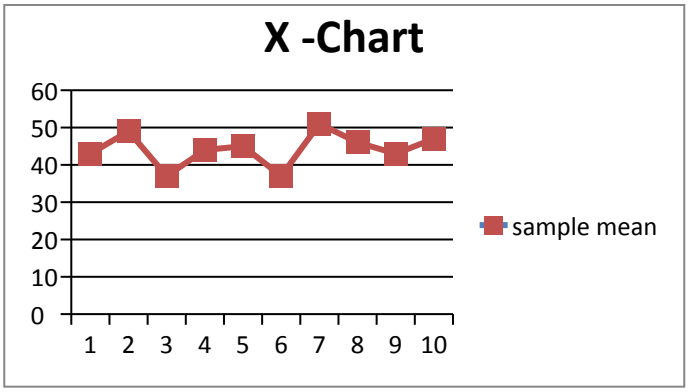
CL(central line) = = 44.2

LCL = - = 44.2-(0.577)(5.8)

= 40.8533440.85

UCL =+ = 44.2+(0.577)(5.8)

=47.5466 47.55



Conclusion:

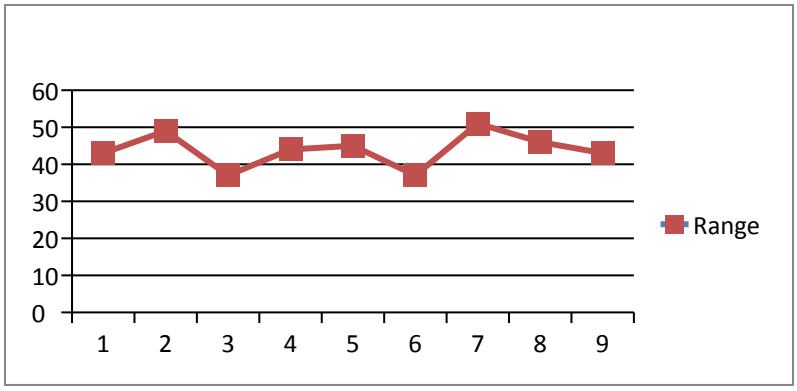
Since 2nd, 3rd, 6th and 7th sample means fall outside the control limits the statistical process is out of control according to chart.

ii) Control limits for R-Chart:

CL = = 5.8, LCL =; =0

UCL = = (2.115) (5.8)

= 12.267 12.27



Conclusion:

Since all the sample mean fall within the control limits the statistical process is under control according to R chart

Inference:

From both and R chart, we see that a point in chart lies outside control limits while all points in R chart lie within control limits ,Through the range variation is under control, we conclude that the process is out of statistical control

Note 1. If the process is to de under control then all sample points in both and R chart must be within control limits

2. Eliminating the sample no.8 which goes outside control limits ,we can get new control limits to set up testing of quality

Example:2

The following are the sample means and range for 10 samples, each of size 5, control chart for mean and range and comment on the nature of control

Sample no	1	2	3	4	5	6	7	8	9	10
Mean	12.8	13.1	13.5	12.9	13.2	14.1	12.1	15.5	13.9	14.2
Range	2.1	3.1	3.9	2.1	1.9	3.0	2.5	2.8	2.5	2.0

Solution:

=

= = 13.53

= = =

= 3.59

From the table of control charts constants for sample size, n=5, = 0.577, = 0, = 2.115

i) Control limits for – Chart:

CL = 13.53

LCL = -

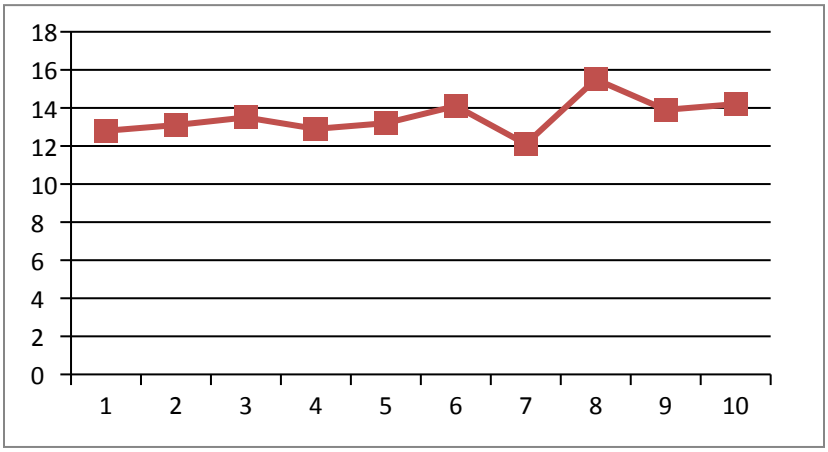
= 13.53 – (0.577) (2.59)

= 12.03557 12.04

UCL = +

= 13.53 + (0.577) (2.59)

= 15.02443 15.02



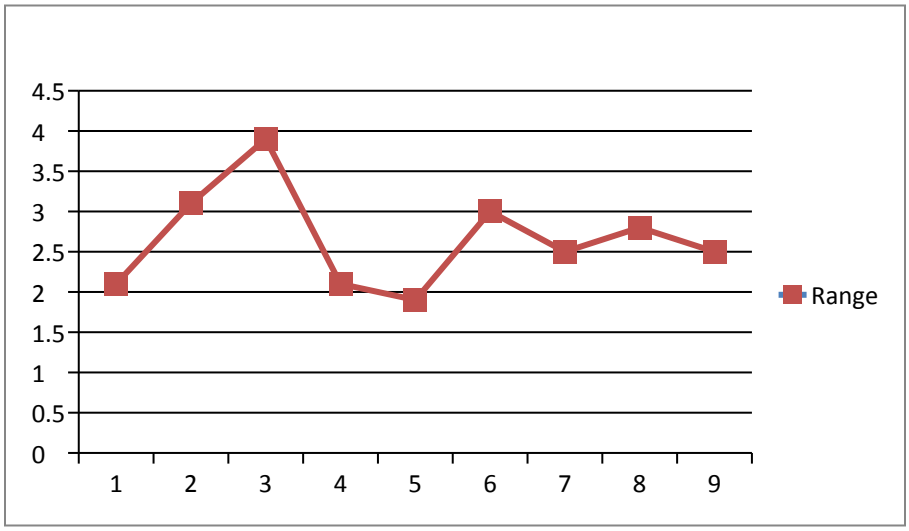
Conclusion: Since 8th sample mean fall outside the control limits the statistical process is out of control according to chart

ii) Control limits for R – Chart:

UCL = = 2.115*2.59 5.48

LCL = = 0

CL = 2.59



Conclusion:

Since all the sample mean fall within the control limits the statistical process is under control according to R chart

Example:3

The following table gives the sample means and ranges for 10 sample each of size 6, in the production of certain component. Construct the control chart for mean and range and comment on the nature of control

Sample no	1	2	3	4	5	6	7	8	9	10
Mean	37.3	49.8	51.5	59.2	54.7	34.7	51.4	61.4	70.7	75.3
Range R	9.5	12.8	10.0	9.1	7.8	5.8	14.5	2.8	3.7	8.0

Sol:

$= = = 54.6$

$= = = = 8.4$

From the table of control chart, for sample size of 6

$= 0.483, = 0, = 2.004$

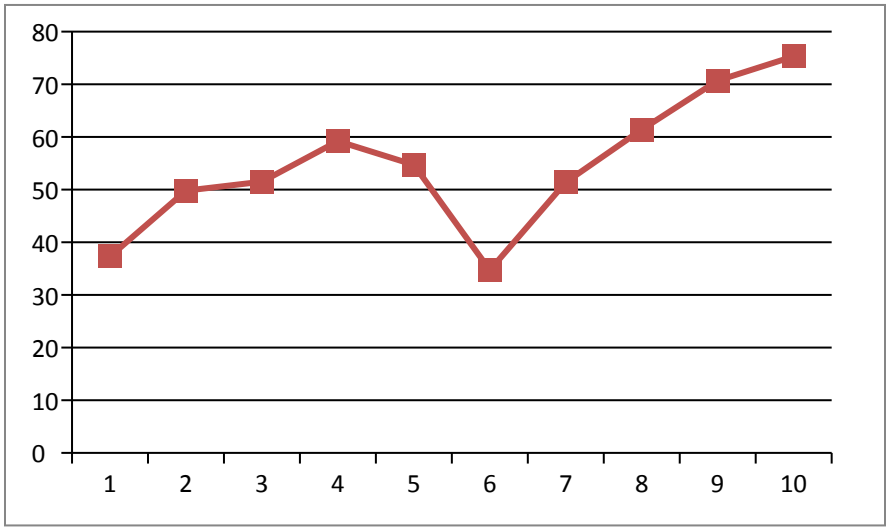
Control limits of Chart:

UCL = +

$= 54.6 + (0.483)\ (8.4) = 58.657$

LCL = -

$= 54.6 – (0.483)\ (8.4) = 50.543$



Conclusion:

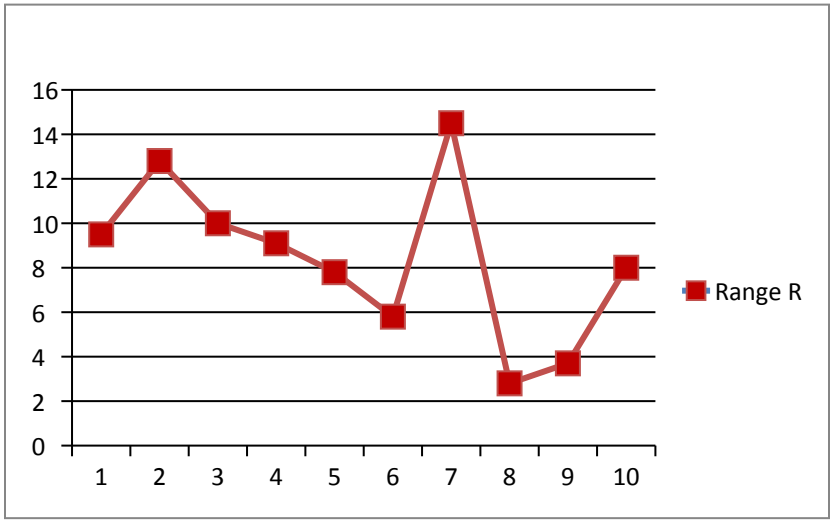
Since 1st ,2nd ,4th ,6th ,8th ,9th, 10th sample means full outside the control limits the statistical process is out of control according to chart,

Control limits of R – Chart:

= 8.4 UCL = = 2.004*8.4

= 16.834

LCL = = 0



Conclusion:

Since all the sample mean fall within the control limits the statistical process is under control according to R chart

Inference;

Though the sample point in R chart lie within control limits, some of the sample points in chart lie outside the control limits. Hence, we conclude that the process is out of control, corrective measures are necessary.

Example:4

The following are the sample means and ranges for ten samples, each of size 5.
Construct the control chart for mean and comment on the nature of control.

Sample No:	1	2	3	4	5	6	7	8	9	10
Mean:	12.8	13.1	13.5	12.9	13.2	14.1	12.1	15.5	13.9	14.2
Range:	2.1	3.1	3.9	2.1	1.9	3	2.5	2.8	2.5	2.0

Solution:

Given size =5 and N=10

$$\overline{X} = \frac{\Sigma X}{N}$$

$$= \frac{12.8 + 13.1 + ... + 14.2}{10} = \frac{135.3}{10}$$

$$= 13.53$$

$$\overline{R} = \frac{\Sigma R_i}{N}$$

$$= \frac{2.1 + 3.1 + ... + 2.0}{10} = \frac{25.9}{10}$$

$$= 2.59$$

From the table of control charts constants for sample size n=5, $A_2 = 0.577, D_3 = 0$ and $D_4 = 2.115$

(i) Control limits for $\overline{X}$ chart:

CL = $\overline{X}$ =13.533

LCL = $\overline{X} - A_2\overline{R}$

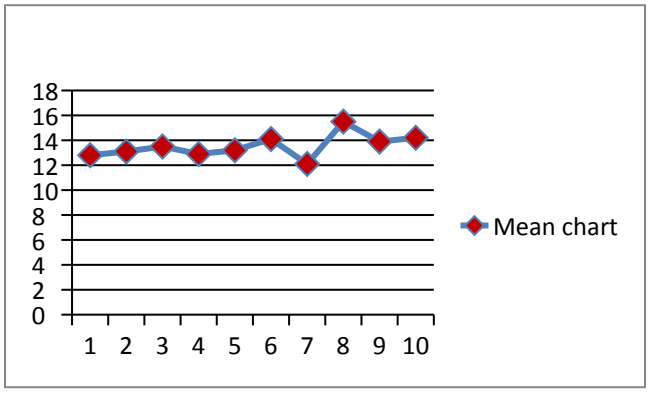
= 13.53-(0.577)(2.59)

= 12.04

UCL = $\overline{X} + A_2 \overline{R}$

= 13.53+(0.577)(2.59)

= 15.02



Conclusion:

Since 8th sample mean fall outside the control limits the statistical process is out of control according to $\overline{X}$ chart.

(i) Control limits for $\overline{R}$ chart:

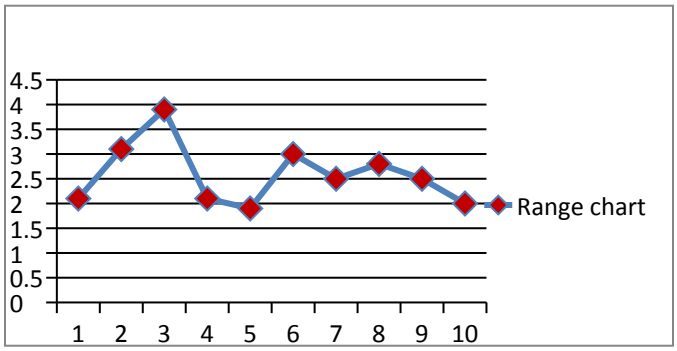
CL = $\overline{R}$ =2.59

LCL = $D_3 \overline{R}$ = 0

UCL = $D_4 \overline{R}$

= 2.115 X 2.59

= 5.48



Conclusion:

Since all the sample mean fall with in the control limits the statistical process is under control

according to $\overline{R}$ chart.

Exercise:

The following are the sample means and range for 10 samples, each of size 5, control chart for mean and range and comment on the nature of control

Sample no	1	2	3	4	5	6	7	8	9	10
Mean	12.8	13.2	13.3	12.7	13.3	14.2	12.2	15.6	13.8	14.3
Range	2.4	3.4	3.8	2.7	1.8	3.1	2.6	2.7	2.4	2.0

Control chart for attributes:

To control the quality of certain products whose attribute are available the following control charts are used

1. np chart of no of defectives
2. p- chart for proportion of defective
3. c- chart for the no of defects in a unit

1) np – Charts:

if the proportion of defectives (successes) in the population of items produced is p. the no. of defectives in a sample of size n is X, then X follows a binomial distribution with means np and S.D is .

Hence p{np- X np+} = 0.9973

The control limits for x, the no. of defectives, are np.

2) p-chart:

Let X follows a normal distribution with mean np and S.D the proportion of defectives follows a normal distribution with mean p and S.D is .

Hence p{p - } = 0.9973.

The control limits for $\frac{X}{n}$, the proportion of the defectives are

As in the case, p is estimated as $\bar{p} = \frac{1}{N}(p_1 + ...p_N)$ where the proportion of defectives in the i^{th} sample is p_i. Hence the control limits for the fraction proportion of sample defectives are $\bar{p} \pm 3\sqrt{\frac{\bar{p}(\sqrt{1-\bar{p}})}{n}}$.

Note:

np- chart and p-chart are used when $\bar{p} \geq .05$ **or** $n\bar{p} \geq 4$.

iii) C-Chart:

It is required to control the no.of defects per unit, c-chart is used, c represents the no.of defects in a unit. For construction of c-chart a record of the no.of defects in each of the N articles inspected should be known since the probability of occurrence of a defect in a unit is very small, the no.of X of defects in a unit follows a Poisson distribution with parameter with mean and S.D $\sqrt{\lambda}$. In the limit X follows a normal distribution with mean and S.D $\sqrt{\lambda}$.

Hence $P\{\lambda - 3\sqrt{\lambda} \leq X \leq \lambda + 3\sqrt{\lambda}\} = 0.9973$

The control limits for X, the no.of defects in a unit are $\lambda \pm 3\sqrt{\lambda}$. Hence the control limits for the no.of defects c in a unit are $\bar{c} + 3\sqrt{c}$.

Example:5

Construct a control chart for defectives for the data:

Sample no:	1	2	3	4	5	6	7	8	9	10
No.of inspected	90	65	85	70	80	80	70	95	90	75
No.of defectives	9	7	3	2	9	5	3	9	6	7

Solution:

we note that the size of the sample varies from sample to sample. We can construct p-chart , provided $0.75\bar{n} < n_i < 1.25\bar{n}$, for all i. Here

$$\bar{n} = \frac{1}{N}\sum n_i = \frac{1}{10}(90 + ... + 75)$$

$$= 80$$

The values of n_i be between 60& 100. Hence p-chart can be drawn by the method given below.

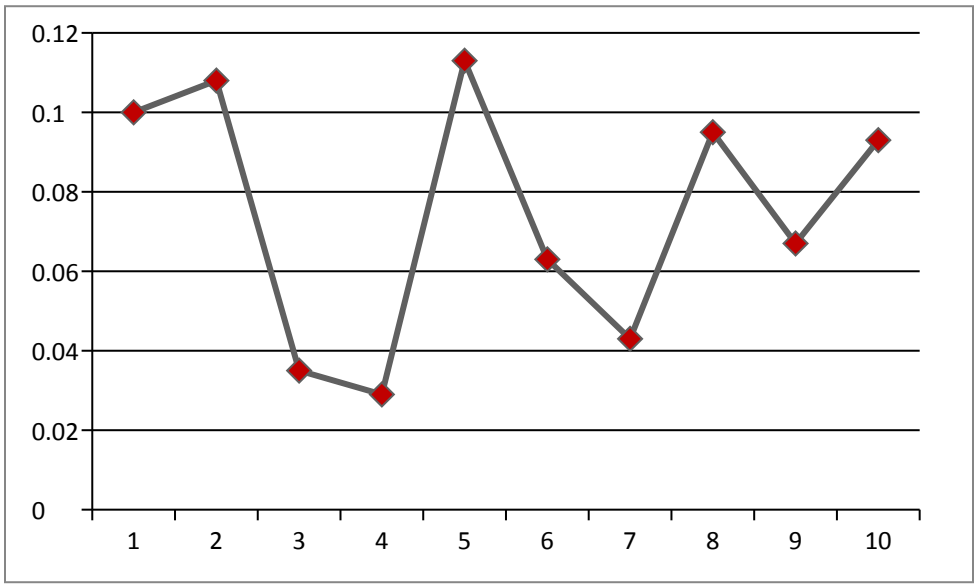
Now, $\bar{p} = \frac{Tota\ln o.ofdefectives}{Tota\ln o.ofitem\sin spected} = 0.075$

Hence for the p-chart to be constructed. CL= $\bar{p} = 0.075$

LCL= $\bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}} = -0.013$

UCL= $\bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}} = 0.0163$

The values of pi for the various samples are 0.100, 0.108,0.035, 0.029, 0.113, 0.063,0.043, .095, 0.067,0.093.



Since all the sample points lie with in the control lines, the process is under control.

Example:6

The data given below are the no. of defectives 10 samples of 100 items each. Construct a p-chart an np-chart and comment on the results:

Sample no:	1	2	3	4	5	6	7	8	9	10
No. of defectives:	6	16	7	3	8	12	7	11	11	4

Solution:

Sample size is constant for all samples n=100

Total no. of defectives = 85

Total no. of inspected = 1000

$$\bar{p} \quad \frac{Tota\ln o.ofdefectives}{Tota\ln o.ofinspected}$$

Average fraction defective = = = 0.085

For p – chart:

LCL = $\bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$

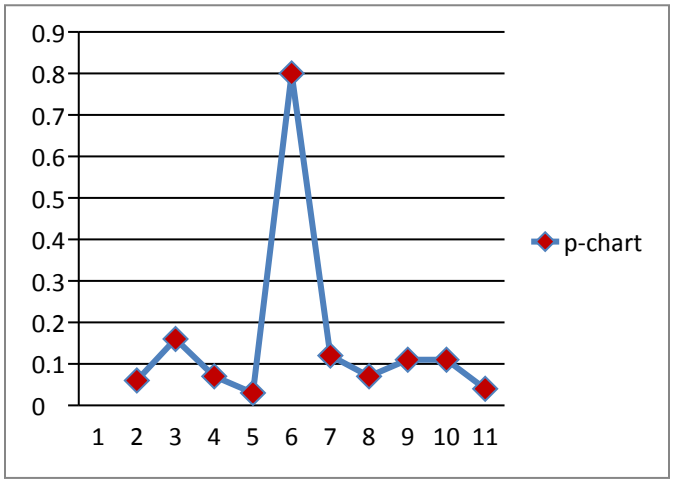
= 0.0013

UCL = $\bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$

= 0.1687

Hence for the p-chart to be constructed. CL= $\bar{p}$ = 0.085

The values of pi for the various samples are 0.06, 0.16, 0.07, 0.03, 0.8, 0.12, 0.07, 0.11, 0.11, 0.04.



Conclusion:

All these values are less than UCL = 0.1687 and LCL = 0.0013. Since all the sample mean fall within the control limits. Hence the process is under statistical control.

For np-chart:

UCL = $n\bar{p} + 3\sqrt{n(1-\bar{p})\bar{p}}$

= 100 x 0.1687

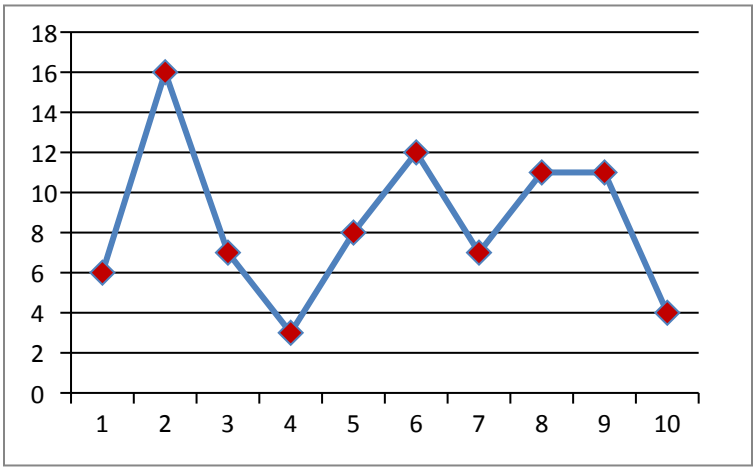
= 16.87

$n\bar{p} = 8.5$

LCL = $n\bar{p} - 3\sqrt{n(1-\bar{p})\bar{p}}$

= 100 x 0.0013

= 0.13



Conclusion:

All these values of no. of defectives in the table lie between 16.87 and 0.13. Since all the sample mean fall within the control limits. Hence the process is under statistical control.

Example:7

A plant produces paper for news print and tolls of paper are inspected for defects.
The results inspections of 20 rolls of papers are given below: Draw the C- chart and comment on the stat of control.

Roll No (i)	1	2	3	4	5	6	7	8	9	10
No. of defects (c):	19	10	8	12	15	22	7	13	18	13
(i)	11	12	13	14	15	16	17	18	19	20
(c):	16	14	8	7	6	4	5	6	8	9

(OR)

Solution:

$$\bar{C} = \frac{1}{N}\Sigma C_i$$

$$= \frac{1}{20} X 220 = 11$$

Control limit $= \bar{C} = 11$

LCL $= \bar{C} - 3\sqrt{\bar{C}}$

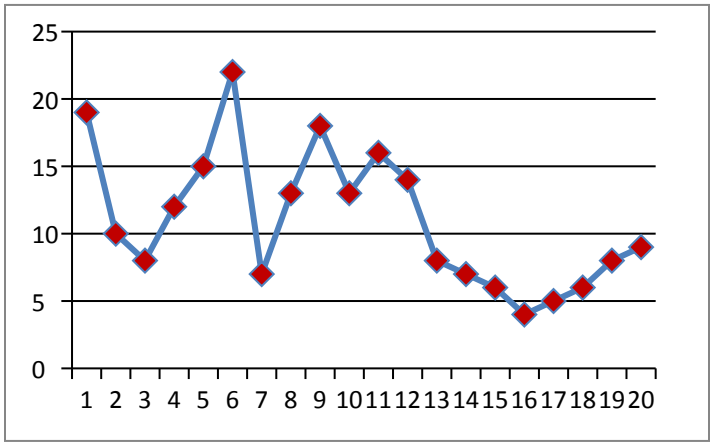
$= 11 - 3\sqrt{11}$

$= 1.05$

UCL $= \bar{C} + 3\sqrt{\bar{C}}$

$= 11 + 3\sqrt{11}$

$= 20.95$



Since one points falls outside the control lines, the process is out of control

Exercise:

1. The following are the figures for the no.of defectives of 10 samples each containing 100 items: 8, 10, 9,8, 10, 11,7, 9,6,12. Draw control chart for fraction defective and comment on the state of control of the process.

2. A plant produces paper for news print and tolls of paper are inspected for defects. The results inspections of 20 rolls of papers are given below: Draw the C- chart and comment on the stat of control.

Roll No (i)	1	2	3	4	5	6	7	8	9	10
No. of defects (c):	19	10	8	12	15	22	7	13	18	13

(i)	11	12	13	14	15	16	17	18	19	20
(c):	16	14	8	7	6	4	5	6	8	9

S.NO	Questions
1	A control chart contains _____ horizontal lines.
2	Attributes are characteristics of products which are _____
3	The theoretical basis for c chart is _____ distribution
4	When the quality of a product is measurable quantitatively, we use control charts are ____
5	The theoretical basis for X chart is _____ distribution
6	Standard error of means _____
7	Variable are those quality characteristics of a product or item which are _____
8	The theoretical basis for the np-chart is ____ distribution.
9	Control chart for number of defects is called _____
10	The total number of defects in 15 pieces of cloth of equal length is 90. Then the UCL for c-chart
11	The variation of a quality characteristics can be divided under _____ heads.
12	The total number of defects in 20 pieces of cloth is 220. The UCL is_____
13	Whenever LCL is ≤ 0, it is taken as _____
14	The total number of defects in 15 pieces of cloth of equal length is 90. Then the UCL
15	Parallel series configuration is also known as _____
16	In R- chart, if σ is known, UCL = D2 σ and LCL = _____
17	The variation of a quality characteristics can be divided under two heads, chance variartion and __
18	Control chart for fraction defective is also called _____
19	Control chart for number of defectives is called _____
20	The theoretical basis for R- chart is _____ distribution
21	The total number of defects in 20 pieces of cloth is 220. The LCL is_____
22	The total number of defects in 15 pieces of cloth of equal length is 90. Then the LCL for c-chart
23	The theoretical basis for c- chart mean is _____
24	For n=2 to 6, the value of D1 is_____
25	In the preparation of R-chart, if D3=0 then LCL is
26	A _____is the partial and total loss of a device
27	_____is only for non repairable items.
28	_____is only for repairable items.
29	The reliability R(t) is_____function of t
30	If the repair time is negligible then MTBF_____
31	Series is _____ in which the components of the system are connected in series
32	Parallel is _____ in which the components of the system are connected in parallel
33	The control limits of R-chart are UCLand LCL_____
34	The theoretical basis for np- chart mean is _____

OPT-1	**OPT-2**	**OPT-3**	**OPT-4**	**OP-5**	**OPT**
UCL	Central line CL	LCL	All		
Normal	Measurablev	Not Measurable	Poission		
Uniform	Binomial	Poisson	Geomentric		
np chart	R chart	X chart	both np &X chart		
Binomial	Poisson	normal	uniform		
A1	A2	A4	A5		
Measurable	Not Measurable	Normal	Poission		
Uniform	Binomial	Poisson	Geomentric		
np chart	R chart	p chart	c chart		
13.35	14.35	0	25.7		
one	two	three	four		
19.95	20.95	0.05	1.05		
0	1	2	3		
13.35	14.35	0	1		
low level redundancy	High level redundan	Redundant configur	Down time		
D1σ	D2 σ	σ	0		
Measurable	Normal	Not Measurable	Assignable variation		
np chart	R chart	p chart	c chart		
np chart	R chart	p chart	c chart		
Binomial	Poisson	normal	uniform		
1.05	0	0.05	1.04		
0	-0.003	0.3	0.001		
λ	np	npq	λp		
2.2	0.1	0.2	0		
0	1	2	3		
failure	reliability	unreliability	availability		
MTTF	MTBF	Hazard rate	Failure		
MTTF	MTBF	Hazard rate	Failure		
normal	increasing	decreasing	failure		
MTTF	MTBF	Hazard rate	MMTF		
one	two	three	zero		
one	two	three	zero		
D3R	D4R	A2R	both D3R&D4R		
np	n/p	p/q	npq		

ANSWERS
All
Not Measurable
Poisson
both np &X chart
normal
A2
Measurable
Binomial
c chart
13.35
two
20.95
0
13.35
low level redundancy
D1σ
Assignable variation
p chart
np chart
normal
1.05
0
λ
0
0
failure
MTTF
MTBF
decreasing
MTTF
one
one
both D3R&D4R
np