QUANTITATIVE TECHNIQUES FOR BUSINESS

COMPLEMENTARY COURSE

BBA (III Semester)

B Com(IV Semester)

(2011 Admission)

UNIVERSITY OF CALICUT

SCHOOL OF DISTANCE EDUCATION

Calicut University P.O. Malappuram, Kerala, India 673 635

412

UNIVERSITY OF CALICUT

SCHOOL OF DISTANCE EDUCATION

STUDY MATERIAL

Complementary Course for

BBA (III Semester)

B Com (IV Semester)

QUANTITATIVE TECHNIQUES FOR BUSINESS

Prepared by	*Sri. Vineethan T,* *Assistant Professor,* *Department of Commerce,* *Govt. College, Madappally.*
Scrutinized by	*Dr. K. Venugopalan,* *Associate Professor,* *Department of Commerce,* *Govt. College, Madappally.*

Layout: *Computer Section, SDE*

CONTENTS

CHAPTER – 1

QUANTITATIVE TECHNIQUES

Meaning and Definition:

Quantitative techniques may be defined as those techniques which provide the decision makes a systematic and powerful means of analysis, based on quantitative data. It is a scientific method employed for problem solving and decision making by the management. With the help of quantitative techniques, the decision maker is able to explore policies for attaining the predetermined objectives. In short, quantitative techniques are inevitable in decision-making process.

Classification of Quantitative Techniques:

There are different types of quantitative techniques. We can classify them into three categories. They are:

1. Mathematical Quantitative Techniques
2. Statistical Quantitative Techniques
3. Programming Quantitative Techniques

Mathematical Quantitative Techcniques:

A technique in which quantitative data are used along with the principles of mathematics is known as mathematical quantitative techniques. Mathematical quantitative techniques involve:

1. Permutations and Combinations:

Permutation means arrangement of objects in a definite order. The number of arrangements depends upon the total number of objects and the number of objects taken at a time for arrangement. The number of permutations or arrangements is calculated by using the following formula:-

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

Combination means selection or grouping objects without considering their order. The number of combinations is calculated by using the following formula:-

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

2. Set Theory:-

Set theory is a modern mathematical device which solves various types of critical problems.

3. Matrix Algebra:

Matrix is an orderly arrangement of certain given numbers or symbols in rows and columns. It is a mathematical device of finding out the results of different types of algebraic operations on the basis of the relevant matrices.

4. Determinants:

It is a powerful device developed over the matrix algebra. This device is used for finding out values of different variables connected with a number of simultaneous equations.

5. Differentiation:

It is a mathematical process of finding out changes in the dependent variable with reference to a small change in the independent variable.

6. Integration:

Integration is the reverse process of differentiation.

7. Differential Equation:

It is a mathematical equation which involves the differential coefficients of the dependent variables.

Statistical Quantitative Techniques:

Statistical techniques are those techniques which are used in conducting the statistical enquiry concerning to certain Phenomenon. They include all the statistical methods beginning from the collection of data till interpretation of those collected data.

Statistical techniques involve:

1. Collection of data:

One of the important statistical methods is collection of data. There are different methods for collecting primary and secondary data.

2. Measures of Central tendency, dispersion, skewness and Kurtosis

Measures of Central tendency is a method used for finding he average of a series while measures of dispersion used for finding out the variability in a series. Measures of Skewness measures asymmetry of a distribution while measures of Kurtosis measures the flatness of peakedness in a distribution.

3. Correlation and Regression Analysis:

Correlation is used to study the degree of relationship among two or more variables. On the other hand, regression technique is used to estimate the value of one variable for a given value of another.

4. Index Numbers:

Index numbers measure the fluctuations in various Phenomena like price, production etc over a period of time, They are described as economic barometres.

5. Time series Analysis:

Analysis of time series helps us to know the effect of factors which are responsible for changes:

6. Interpolation and Extrapolation:

Interpolation is the statistical technique of estimating under certain assumptions, the missing figures which may fall within the range of given figures. Extrapolation provides estimated figures outside the range of given data.

7. Statistical Quality Control

Statistical quality control is used for ensuring the quality of items manufactured. The variations in quality because of assignable causes and chance causes can be known with the help of this tool. Different control charts are used in controlling the quality of products.

8. Ratio Analysis:

Ratio analysis is used for analyzing financial statements of any business or industrial concerns which help to take appropriate decisions.

9. Probability Theory:

Theory of probability provides numerical values of the likely hood of the occurrence of events.

10. Testing of Hypothesis

Testing of hypothesis is an important statistical tool to judge the reliability of inferences drawn on the basis of sample studies.

Programming Techniques:

Programming techniques are also called operations research techniques. Programming techniques are model building techniques used by decision makers in modern times.

Programming techniques involve:

1. **Linear Programming**:

Linear programming technique is used in finding a solution for optimizing a given objective under certain constraints.

2. **Queuing Theory**:

Queuing theory deals with mathematical study of queues. It aims at minimizing cost of both servicing and waiting.

3. **Game Theory**:

Game theory is used to determine the optimum strategy in a competitive situation.

4. **Decision Theory**:

This is concerned with making sound decisions under conditions of certainty, risk and uncertainty.

5. **Inventory Theory**:

Inventory theory helps for optimizing the inventory levels. It focuses on minimizing cost associated with holding of inventories.

6. **Net work programming**:

It is a technique of planning, scheduling, controlling, monitoring and co-ordinating large and complex projects comprising of a number of activities and events. It serves as an instrument in resource allocation and adjustment of time and cost up to the optimum level. It includes CPM, PERT etc.

7. **Simulation**:

It is a technique of testing a model which resembles a real life situations

8. **Replacement Theory**:

It is concerned with the problems of replacement of machines, etc due to their deteriorating efficiency or breakdown. It helps to determine the most economic replacement policy.

9. Non Linear Programming:

It is a programming technique which involves finding an optimum solution to a problem in which some or all variables are non-linear.

10. **Sequencing**:

Sequencing tool is used to determine a sequence in which given jobs should be performed by minimizing the total efforts.

11. **Quadratic Programming**:

Quadratic programming technique is designed to solve certain problems, the objective function of which takes the form of a quadratic equation.

12. Branch and Bound Technique

It is a recently developed technique. This is designed to solve the combinational problems of decision making where there are large number of feasible solutions. Problems of plant location, problems of determining minimum cost of production etc. are examples of combinational problems.

Functions of Quantitative Techniques:

The following are the important functions of quantitative techniques:

1. To facilitate the decision-making process
2. To provide tools for scientific research
3. To help in choosing an optimal strategy
4. To enable in proper deployment of resources
5. To help in minimizing costs
6. To help in minimizing the total processing time required for performing a set of jobs

USES OF QUANTITATE TECHNIQUES

Business and Industry

Quantitative techniques render valuable services in the field of business and industry. Today, all decisions in business and industry are made with the help of quantitative techniques.

Some important uses of quantitative techniques in the field of business and industry are given below:

1. Quantitative techniques of linear programming is used for optimal allocation of scarce resources in the problem of determining product mix

2. Inventory control techniques are useful in dividing when and how much items are to be purchase so as to maintain a balance between the cost of holding and cost of ordering the inventory

3. Quantitative techniques of CPM, and PERT helps in determining the earliest and the latest times for the events and activities of a project. This helps the management in proper deployment of resources.

4. Decision tree analysis and simulation technique help the management in taking the best possible course of action under the conditions of risks and uncertainty.

5. Queuing theory is used to minimize the cost of waiting and servicing of the customers in queues.

6. Replacement theory helps the management in determining the most economic replacement policy regarding replacement of an equipment.

Limitations of Quantitative Techniques:

Even though the quantitative techniques are inevitable in decision-making process, they are not free from short comings. The following are the important limitations of quantitative techniques:

1. Quantitative techniques involves mathematical models, equations and other mathematical expressions
2. Quantitative techniques are based on number of assumptions. Therefore, due care must be ensured while using quantitative techniques, otherwise it will lead to wrong conclusions.
3. Quantitative techniques are very expensive.
4. Quantitative techniques do not take into consideration intangible facts like skill, attitude etc.
5. Quantitative techniques are only tools for analysis and decision-making. They are not decisions itself.

CHAPTER - 2

CORRELEATION ANALYSIS

Introduction:

In practice, we may come across with lot of situations which need statistical analysis of either one or more variables. The data concerned with one variable only is called univariate data. For Example: Price, income, demand, production, weight, height marks etc are concerned with one variable only. The analysis of such data is called univariate analysis.

The data concerned with two variables are called bivariate data. For example: rainfall and agriculture; income and consumption; price and demand; height and weight etc. The analysis of these two sets of data is called bivariate analysis.

The date concerned with three or more variables are called multivariate date. For example: agricultural production is influenced by rainfall, quality of soil, fertilizer etc.

The statistical technique which can be used to study the relationship between two or more variables is called correlation analysis.

Definition:

Two or more variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.

According to Simpson and Kafka, "Correlation analysis deals with the association between two or more variables".

Lun chou defines, " Correlation analysis attempts to determine the degree of relationship between variables".

Boddington states that "Whenever some definite connection exists between two or more groups or classes of series of data, there is said to be correlation."

In nut shell, correlation analysis is an analysis which helps to determine the degree of relationship exists between two or more variables.

Correlation Coefficient:

Correlation analysis is actually an attempt to find a numerical value to express the extent of relationship exists between two or more variables. The numerical measurement showing the degree of correlation between two or more variables is called correlation coefficient. Correlation coefficient ranges between -1 and +1.

SIGNIFICANCE OF CORRELATION ANALYSIS

Correlation analysis is of immense use in practical life because of the following reasons:

1. Correlation analysis helps us to find a single figure to measure the degree of relationship exists between the variables.
2. Correlation analysis helps to understand the economic behavior.

3. Correlation analysis enables the business executives to estimate cost, price and other variables.
4. Correlation analysis can be used as a basis for the study of regression. Once we know that two variables are closely related, we can estimate the value of one variable if the value of other is known.
5. Correlation analysis helps to reduce the range of uncertainty associated with decision making. The prediction based on correlation analysis is always near to reality.
6. It helps to know whether the correlation is significant or not. This is possible by comparing the correlation co-efficient with 6PE. It 'r' is more than 6 PE, the correlation is significant.

Classification of Correlation

Correlation can be classified in different ways. The following are the most important classifications

1. Positive and Negative correlation
2. Simple, partial and multiple correlation
3. Linear and Non-linear correlation

Positive and Negative Correlation

Positive Correlation

When the variables are varying in the same direction, it is called positive correlation. In other words, if an increase in the value of one variable is accompanied by an increase in the value of other variable or if a decrease in the value of one variable is accompanied by a decree se in the value of other variable, it is called positive correlation.

Eg: 1)	A:	10	20	30	40	50
	B:	80	100	150	170	200
2)	X:	78	60	52	46	38
	Y:	20	18	14	10	5

Negative Correlation:

When the variables are moving in opposite direction, it is called negative correlation. In other words, if an increase in the value of one variable is accompanied by a decrease in the value of other variable or if a decrease in the value of one variable is accompanied by an increase in the value of other variable, it is called negative correlation.

Eg: 1)	A:	5	10	15	20	25
	B:	16	10	8	6	2

2) X: 40 32 25 20 10

Y: 2 3 5 8 12

Simple, Partial and Multiple correlation

Simple Correlation

In a correlation analysis, if only two variables are studied it is called simple correlation. Eg. the study of the relationship between price & demand, of a product or price and supply of a product is a problem of simple correlation.

Multiple correlation

In a correlation analysis, if three or more variables are studied simultaneously, it is called multiple correlation. For example, when we study the relationship between the yield of rice with both rainfall and fertilizer together, it is a problem of multiple correlation.

Partial correlation

In a correlation analysis, we recognize more than two variable, but consider one dependent variable and one independent variable and keeping the other Independent variables as constant. For example yield of rice is influenced b the amount of rainfall and the amount of fertilizer used. But if we study the correlation between yield of rice and the amount of rainfall by keeping the amount of fertilizers used as constant, it is a problem of partial correlation.

Linear and Non-linear correlation

Linear Correlation

In a correlation analysis, if the ratio of change between the two sets of variables is same, then it is called linear correlation.

For example when 10% increase in one variable is accompanied by 10% increase in the other variable, it is the problem of linear correlation.

X: 10 15 30 60

Y: 50 75 150 300

Here the ratio of change between X and Y is the same. When we plot the data in graph paper, all the plotted points would fall on a straight line.

Non-linear correlation

In a correlation analysis if the amount of change in one variable does not bring the same ratio of change in the other variable, it is called non linear correlation.

X: 2 4 6 10 15

Y: 8 10 18 22 26

Here the change in the value of X does not being the same proportionate change in the value of Y.

This is the problem of non-linear correlation, when we plot the data on a graph paper, the plotted points would not fall on a straight line.

Degrees of correlation:

Correlation exists in various degrees

1. **Perfect positive correlation**

If an increase in the value of one variable is followed by the same proportion of increase in other related variable or if a decrease in the value of one variable is followed by the same proportion of decrease in other related variable, it is perfect positive correlation. eg: if 10% rise in price of a commodity results in 10% rise in its supply, the correlation is perfectly positive. Similarly, if 5% full in price results in 5% fall in supply, the correlation is perfectly positive.

2. **Perfect Negative correlation**

If an increase in the value of one variable is followed by the same proportion of decrease in other related variable or if a decrease in the value of one variable is followed by the same proportion of increase in other related variably it is Perfect Negative Correlation. For example if 10% rise in price results in 10% fall in its demand the correlation is perfectly negative. Similarly if 5% fall in price results in 5% increase in demand, the correlation is perfectly negative.

3. **Limited Degree of Positive correlation**:

When an increase in the value of one variable is followed by a non-proportional increase in other related variable, or when a decrease in the value of one variable is followed by a non-proportional decrease in other related variable, it is called limited degree of positive correlation.

For example, if 10% rise in price of a commodity results in 5% rise in its supply, it is limited degree of positive correlation. Similarly if 10% fall in price of a commodity results in 5% fall in its supply, it is limited degree of positive correlation.

4. **Limited degree of Negative correlation**

When an increase in the value of one variable is followed by a non-proportional decrease in other related variable, or when a decrease in the value of one variable is followed by a non-proportional increase in other related variable, it is called limited degree of negative correlation.

For example, if 10% rise in price results in 5% fall in its demand, it is limited degree of negative correlation. Similarly, if 5% fall in price results in 10% increase in demand, it is limited degree of negative correlation.

5. **Zero Correlation (Zero Degree correlation)**

If there is no correlation between variables it is called zero correlation. In other words, if the values of one variable cannot be associated with the values of the other variable, it is zero correlation.

Methods of measuring correlation

Correlation between 2 variables can be measured by graphic methods and algebraic methods.

I Graphic Methods

1) Scatter Diagram

2) Correlation graph

II Algebraic methods (Mathematical methods or statistical methods or Co-efficient of correlation methods):

1) Karl Pearson's Co-efficient of correlation

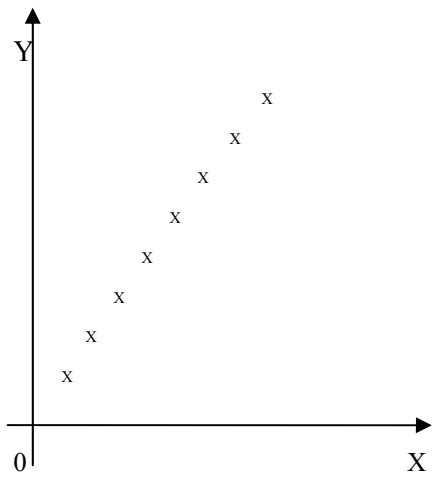
2) Spear mans Rank correlation method

3) Concurrent deviation method

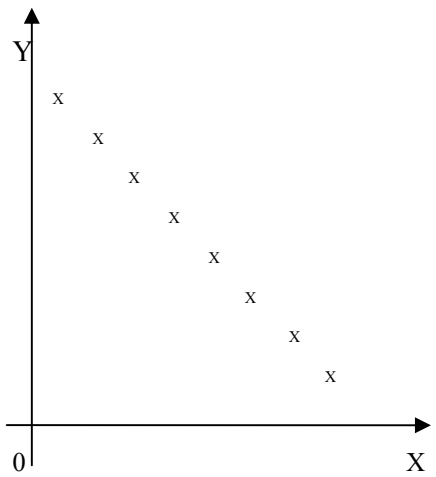
Scatter Diagram

This is the simplest method for ascertaining the correlation between variables. Under this method all the values of the two variable are plotted in a chart in the form of dots. Therefore, it is also known as dot chart. By observing the scatter of the various dots, we can form an idea that whether the variables are related or not.

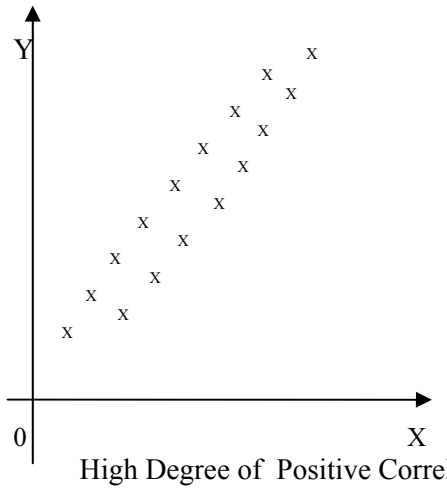
A scatter diagram indicates the direction of correlation and tells us how closely the two variables under study are related. The greater the scatter of the dots, the lower is the relationship



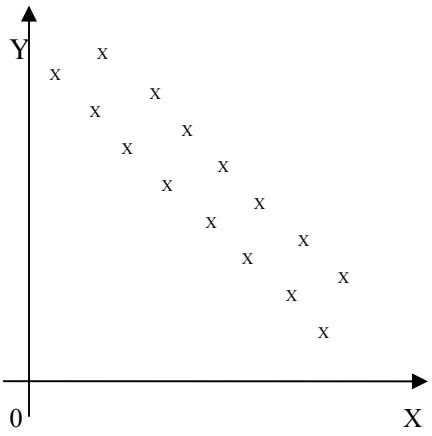
Perfect Positive Correlation



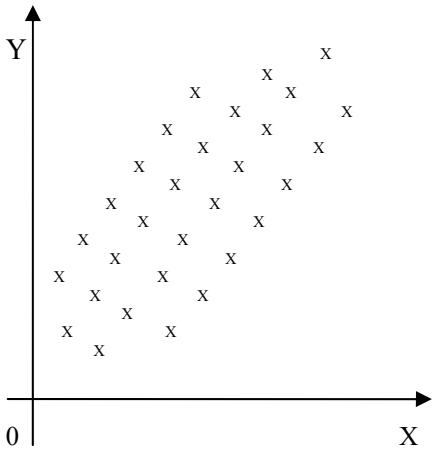
Perfect Negative Correlation



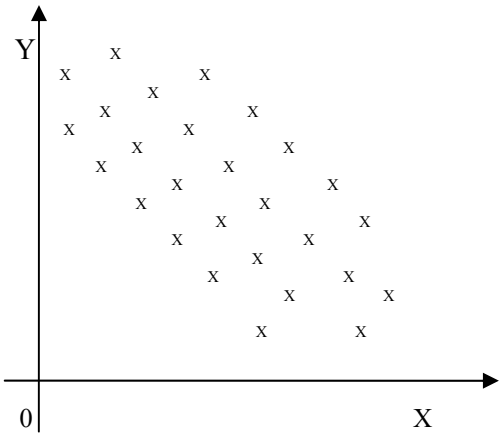
High Degree of Positive Correlation



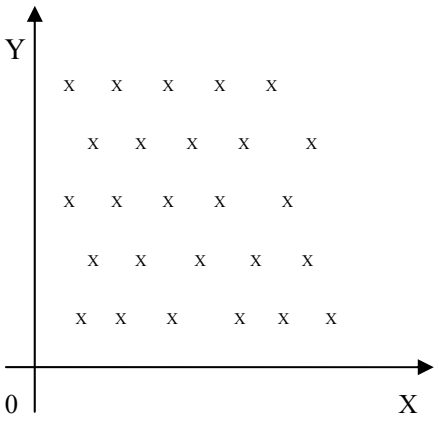
High Degree of Negative Correlation



Low Degree of Positive Correlation



Low Degree of Negative Correlation



No Correlation (r = 0)

Merits of Scatter Diagram method

1. It is a simple method of studying correlation between variables.
2. It is a non-mathematical method of studying correlation between the variables. It does not require any mathematical calculations.
3. It is very easy to understand. It gives an idea about the correlation between variables even to a layman.
4. It is not influenced by the size of extreme items.
5. Making a scatter diagram is, usually, the first step in investigating the relationship between two variables.

Demerits of Scatter diagram method

1. It gives only a rough idea about the correlation between variables.
2. The numerical measurement of correlation co-efficient cannot be calculated under this method.
3. It is not possible to establish the exact degree of relationship between the variables.

Correlation graph Method

Under correlation graph method the individual values of the two variables are plotted on a graph paper. Then dots relating to these variables are joined separately so as to get two curves. By examining the direction and closeness of the two curves, we can infer whether the variables are related or not. If both the curves are moving in the same direction(either upward or downward) correlation is said to be positive. If the curves are moving in the opposite directions, correlation is said to be negative.

Merits of Correlation Graph Method

1. This is a simple method of studying relationship between the variable
2. This does not require mathematical calculations.
3. This method is very easy to understand

Demerits of correlation graph method:

1. A numerical value of correlation cannot be calculated.
2. It is only a pictorial presentation of the relationship between variables.
3. It is not possible to establish the exact degree of relationship between the variables.

Karl Pearson's Co-efficient of Correlation

Karl Pearson's Coefficient of Correlation is the most popular method among the algebraic methods for measuring correlation. This method was developed by Prof. Karl Pearson in 1896. It is also called product moment correlation coefficient.

Pearson's coefficient of correlation is defined as the ratio of the covariance between X and Y to the product of their standard deviations. This is denoted by 'r' or r_{xy}

$$r = \frac{\text{Covariance of X and Y}}{(\text{SD of X}) \times (\text{SD of Y})}$$

Interpretation of Co-efficient of Correlation

Pearson's Co-efficient of correlation always lies between +1 and -1. The following general rules will help to interpret the Co-efficient of correlation:

1. When r - +1, It means there is perfect positive relationship between variables.
2. When r = -1, it means there is perfect negative relationship between variables.
3. When r = 0, it means there is no relationship between the variables.
4. When 'r' is closer to +1, it means there is high degree of positive correlation between variables.
5. When 'r' is closer to – 1, it means there is high degree of negative correlation between variables.
6. When 'r' is closer to 'O', it means there is less relationship between variables.

Properties of Pearson's Co-efficient of Correlation

1. If there is correlation between variables, the Co-efficient of correlation lies between +1 and -1.
2. If there is no correlation, the coefficient of correlation is denoted by zero (ie r=0)
3. It measures the degree and direction of change
4. If simply measures the correlation and does not help to predict cansation.
5. It is the geometric mean of two regression co-efficients.

i.e $r = \sqrt{b_{xy} \cdot b_{yx}}$

Computation of Pearson's Co-efficient of correlation:

Pearson's correlation co-efficient can be computed in different ways. They are:

a Arithmetic mean method

b Assumed mean method

c Direct method

Arithmetic mean method:-

Under arithmetic mean method, co-efficient of correlation is calculated by taking actual mean.

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}$$

or

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \in y^2}}$$ whereas x-x-$\bar{x}$ and y=y-$\bar{y}$

Calculate Pearson's co-efficient of correlation between age and playing habits of students:

Age:	20	21	22	23	24	25
No. of students	500	400	300	240	200	160
Regular players	400	300	180	96	60	24

Let X = Age and Y = Percentage of regular players

Percentage of regular players can be calculated as follows:-

$\frac{400}{500}$ x 100 = 80; $\frac{300}{400}$ x 100 = 75; $\frac{180}{300}$ x 100 = 60; $\frac{96}{240}$ x 100 = 40 ,

$\frac{60}{20}$ x 100 = 30; and $\frac{24}{160}$ x 100 = 15

Pearson's Coefficient of Correlation (r) $= \frac{\sum(x-\bar{x}).(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2.(y-\bar{y})^2}}$

Computation of Pearson's Coefficient of correlation

Age x	% of Regular Player y	$x-\bar{x}$ (x-22.5)	$(y-\bar{y})$ (y-50)	$(x-\bar{x})\,(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
20	80	-2.5	30	-75.0	6.25	900
21	75	-1.5	25	-37.5	2.25	625
22	60	-0.5	10	- 5.0	0.25	100
23	40	0.5	-10	- 5.0	0.25	100
24	30	1.5	-20	-30.0	2.25	400
25	15	2.5	-35	-87.5	6.25	1225
135	300			-240	17.50	3350

$$\bar{x} = \frac{\Sigma x}{N} = \frac{135}{6} = 22.5$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{300}{6} = 50$$

$$r = \frac{240}{\sqrt{17.5 \times 3350}} = \frac{-240}{\sqrt{58,625}} = \frac{-240}{\sqrt{242.126}} = -0.9912$$

Assumed mean method:

Under assumed mean method, correlation coefficient is calculated by taking assumean only.

$$r = \frac{N\,\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2} \times \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

Where dx = deviations of X from its assumed mean; dy= deviations of y from its assumed mean

Find out coefficient of correlation between size and defect in quality of shoes:

Size	:	15-16	16-17	17-18	18-19	19-20	20-21
No. of shoes Produced	:	200	270	340	360	400	300
No. of defectives:		150	162	170	180	180	114

Let x = size (ie mid-values)

y = percentage of defectives

∴ x values are 15.5 , 16.5, 17.5, 18.5, 19.5 and 20.5

y values are 75, 60, 50, 50, 45 and 38

Take assumed mean: x = 17.5 and y = 50

Computation of Pearson's Coefficient of Correlation						
x	y	dx	dy	dxdy	dx^2	dy^2
15.5	75	-2	25	-50	4	625
16.5	60	-1	10	-10	1	100
17.5	50	0	0	0	0	0
18.5	50	1	0	0	1	0
19.5	45	2	-5	-10	4	25
20.5	38	3	-12	-36	9	144
		$\Sigma dx = 3$	$\Sigma dy = 18$	$\Sigma dxdy = -106$	$\Sigma dx^2 = 19$	$\Sigma dy^2 = 894$

$$r = \frac{N\,\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{\sqrt{N\Sigma dx^{2-}(\Sigma dx)^2} \times \sqrt{N\Sigma dy^{2-}(\Sigma dy)^2}}$$

$$r = \frac{(6x - 106) - (3x18)}{\sqrt{(6\times19) - 3^2} \times \sqrt{(6\times894) - 18^2}}$$

$$= \frac{-636 - 54}{\sqrt{114 - 9} \times \sqrt{5364 - 324}}$$

$$= \frac{-690}{\sqrt{105} \times \sqrt{5040}} = \frac{-690}{727.46} = -0.9485$$

Direct Method:

Under direct method, coefficient of correlation is calculated without taking actual mean or assumed mean

$$r = \frac{N\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{N\Sigma x^{2-}(\Sigma x)^2} \times \sqrt{N\Sigma y^{2-}(\Sigma y)^2}}$$

From the following data, compute Pearson's correlation coefficient:

Price : 10 12 14 15 19

Demand (Qty) 40 41 48 60 50

Let us take price = x and demand = y

Computation of Pearson's Coefficient of Correlation				
Price (x)	**Demand (y)**	**xy**	**x^2**	**y^2**
10	40	400	100	1600
12	41	492	144	1681
14	48	672	196	2304
15	60	900	225	3600
19	50	950	361	2500
$\Sigma x = 70$	$\Sigma y = 239$	$\Sigma xy = 3414$	$\Sigma x^2 = 1026$	$\Sigma y^2 = 11685$

$$r = \frac{N\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{N\Sigma x^2 - (\Sigma x)^2} \times \sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$$r = \frac{(5 \times 3414) - (70 \times 239)}{\sqrt{(5x1026)^{-}70^{2\times}}\sqrt{(5 \times 11685) - 239^2}}$$

$$r = \frac{17{,}070 - 16{,}730}{\sqrt{230} \times \sqrt{1304}} = \frac{340}{547.65} = {}^{+}0.621$$

Probable Error and Coefficient of Correlation

Probable error (PE) of the Co-efficient of correlation is a statistical device which measures the reliability and dependability of the value of co-efficient of correlation.

Probable Error $= \frac{2}{3}$ standard error

= 0.6745 x standard error

Standard Error (SE) $= \frac{1 - r^2}{\sqrt{n}}$

$$\therefore PE = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

If the value of coefficient of correlation (r) is less than the PE, then there is no evidence of correlation.

If the value of 'r' is more than 6 times of PE, the correlation is certain and significant.

By adding and submitting PE from coefficient of correlation, we can find out the upper and lower limits within which the population coefficient of correlation may be expected to lie.

Uses of PE:

1) PE is used to determine the limits within which the population coefficient of correlation may be expected to lie.

2) It can be used to test whether the value of correlation coefficient of a sample is significant with that of the population

If r = 0.6 and N = 64, find out the PE and SE of the correlation coefficient. Also determine the limits of population correlation coefficient.

Sol: r = 0.6

N=64

PE = 0.6745×SE

SE = $\frac{1-r^2}{\sqrt{n}}$

= $\frac{1-(0.6)^2}{\sqrt{64}} = \frac{1-0.36}{8} = \frac{0.64}{8} = \underline{0.08}$

P.E = 0.6745 × 0.08

= $\underline{\underline{0.05396}}$

Limits of population Correlation coefficient = r ± PE

= 0.6 ±0.05396

= $\underline{\underline{0.54604 \text{ to } 0.6540}}$

Qn. 2 r and PE have values 0.9 and 0.04 for two series. Find n.

Sol: PE = 0.04

$0.6745 \times \frac{1-r^2}{\sqrt{n}} = 0.04$

$\frac{1-0.9^2}{\sqrt{n}} = \frac{0.04}{0.6745}$

$\frac{1-0.81}{\sqrt{n}} = 0.0593$

$\frac{0.19}{\sqrt{n}} = 0.0593$

$0.0593 \times \sqrt{n} = 0.19$

$\sqrt{n} = \frac{0.19}{0.0593}$

$\sqrt{n} = 3.2$

$N = 3.2^2 = 10.266$

$\underline{\underline{N = 10}}$

Coefficient of Determination

One very convenient and useful way of interpreting the value of coefficient of correlation is the use of the square of coefficient of correlation. The square of coefficient of correlation is called coefficient of determination.

Coefficient of determination = r^2

Coefficient of determination is the ratio of the explained variance to the total variance.

For example, suppose the value of r = 0.9, then r^2 = 0.81=81%

This means that 81% of the variation in the dependent variable has been explained by (determined by) the independent variable. Here 19% of the variation in the dependent variable has not been explained by the independent variable. Therefore, this 19% is called coefficient of non-determination.

Coefficient of non-determination (K^2) = $1 – r^2$

K^2 = 1- coefficient of determination

Qn: Calculate coefficient of determination and non-determination if coefficient of correlation is 0.8

Sol:- r =0.8

Coefficient of determination $= r^2$

$= 0.8^2 = 0.64 = 64\%$

Co efficient of non-determination $= 1 – r^2$

$= 1 - 0.64$

$= 0.36$

$= \underline{\underline{36\%}}$

Merits of Pearson's Coefficient of Correlation:-

1. This is the most widely used algebraic method to measure coefficient of correlation.
2. It gives a numerical value to express the relationship between variables
3. It gives both direction and degree of relationship between variables
4. It can be used for further algebraic treatment such as coefficient of determination coefficient of non-determination etc.
5. It gives a single figure to explain the accurate degree of correlation between two variables

Demerits of Pearson's Coefficient of correlation

1. It is very difficult to compute the value of coefficient of correlation.
2. It is very difficult to understand

3. It requires complicated mathematical calculations
4. It takes more time
5. It is unduly affected by extreme items
6. It assumes a linear relationship between the variables. But in real life situation, it may not be so.

Spearman's Rank Correlation Method

Pearson's coefficient of correlation method is applicable when variables are measured in quantitative form. But there were many cases where measurement is not possible because of the qualitative nature of the variable. For example, we cannot measure the beauty, morality, intelligence, honesty etc in quantitative terms. However it is possible to rank these qualitative characteristics in some order.

The correlation coefficient obtained from ranks of the variables instead of their quantitative measurement is called rank correlation. This was developed by Charles Edward Spearman in 1904.

Spearman's coefficient correlation (R) = $1- \frac{6\Sigma D^2}{N^{3-N}}$

Where D = difference of ranks between the two variables

N = number of pairs

Qn: Find the rank correlation coefficient between poverty and overcrowding from the information given below:

Town:	A	B	C	D	E	F	G	H	I	J
Poverty:	17	13	15	16	6	11	14	9	7	12
Over crowing:	36	46	35	24	12	18	27	22	2	8

Sol: Here ranks are not given. Hence we have to assign ranks

$R = 1- \frac{6\Sigma D^2}{N^{3-N}}$

N = 10

Computation of rank correlation Co-efficient						
Town	Poverty	Over crowding	R_1	R_2	D	D^2
A	17	36	1	2	1	1
B	13	46	5	1	4	16
C	15	35	3	3	0	0
D	16	24	2	5	3	9
E	6	12	10	8	2	4
F	11	18	7	7	0	0
G	14	27	4	4	0	0
H	9	22	8	6	2	4
I	7	2	9	10	1	1
J	12	8	6	9	3	9
					ΣD^2	44

$$R = 1 - \frac{6 \times 44}{10^3 - 10}$$

$$= 1 - \frac{264}{990}$$

$$= 1 - 0.2667$$

$$= {}^{+}\underline{0.7333}$$

Qn:- Following were the ranks given by three judges in a beauty context. Determine which pair of judges has the nearest approach to Common tastes in beauty.

Judge I:	1	6	5	10	3	2	4	9	7	8
Judge I:	3	5	8	4	7	10	2	1	6	9
Judge I:	6	4	9	8	1	2	3	10	5	7

$$R = 1 = \frac{6\Sigma D^2}{N^3 - N}$$

N= 10

Computation of Spearman's Rank Correlation Coefficient								
Judge I (R_1)	Judge II (R_2)	Judge III (R_3)	R_1-R_2 (D_1)	R_2-R_3 (D_2)	R_1-R_3 (D_3)	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
				ΣD^2		200	214	60

$$R = 1 - \frac{6\Sigma D^2}{N}$$

Rank correlation coefficient between I & II = $\frac{6\times200}{10^3-10}$

$= 1 - \frac{1200}{990}$

$= 1 - 1.2121$

$= -\underline{0.2121}$

Rank correlation Coefficient between II & III judges $= 1 - \frac{6\times214}{10^3-10}$

$= 1 - \frac{1284}{990}$

$= -\underline{0.297}$

Rank correlation coefficient between I& II judges $= 1 - \frac{6\times60}{10^3-10}$

$= 1 - \frac{360}{990}$

$= 1 - 0.364$

$= +\underline{0.636}$

The rank correlation coefficient in case of I & III judges is greater than the other two pairs. Therefore, judges I & III have highest similarity of thought and have the nearest approach to common taste in beauty.

Qn: The Co-efficient of rank correlation of the marks obtained by 10 students in statistics & English was 0.2. It was later discovered that the difference in ranks of one of the students was wrongly takes as 7 instead of 9 Find the correct result.

$R = 0.2$

$R = 1 - \frac{6\Sigma D^2}{N3-N} = 0.2$

$\frac{1-0.2}{1} = \frac{6\Sigma D^2}{10^3-10}$

$\frac{0.8}{1} = \frac{6\Sigma D^2}{990}$

$6\Sigma D^2 = 90 \times 0.8 = 792$

Correct $\Sigma D^2 = \frac{792}{6} = 132 - 7^2 + 9^2$

$= \underline{164}$

Correct R $1 = \frac{6\Sigma D^2}{N^3-N}$

$= 1 - \frac{6\times164}{10^3-10}$

$= 1 - \frac{984}{990}$

$= 1 - 0.9939$

$= \underline{0.0061}$

Qn: The coefficient of rank correlation between marks in English and maths obtained by a group students is 0.8. If the sum of the squares of the difference in ranks is given to be 33, find the number of students in the group.

Sol: $R = 1 - \frac{6\Sigma D^2}{N^3-N} = 0.8$

ie, $1 - \frac{6 \times 33}{N^3-N} = 0.8$

$1-08 = \frac{6 \times 33}{N^3-N}$

$0.2 \times (N^3 - N) = 198$

$N^3 - N = \frac{198}{0.2} = 990$

N = 10

Computation of Rank Correlation Coefficient when Ranks are Equal

There may be chances of obtaining same rank for two or more items. In such a situation, it is required to give average rank for all. Such items. For example, if two observations got 4th rank, each of those observations should be given the rank 4.5 (i.e $\frac{4+5}{2} = 4.5$)

Suppose 4 observations got 6th rank, here we have to assign the rank, 7.5 (ie. $\frac{6+7+8+9}{4}$) to each of the 4 observations.

When there is equal ranks, we have to apply the following formula to compute rank correlation coefficient:-

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

Where D – Difference of rank in the two series

N - Total number of pairs

m - Number of times each rank repeats

Qn:- Obtain rank correlation co-efficient for the data:-

X :	68	64	75	50	64	80	75	40	55	64
Y:	62	58	68	45	81	60	68	48	50	70

Here, ranks are not given we have to assign ranks Further, this is the case of equal ranks.

$$\therefore R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

Computation of rank correlation coefficient					
x	y	R_1	R_2	$D(R_1-R_2)$	D^2
68	62	4	5	1	1
64	58	6	7	1	1
75	68	2.5	3.5	1	1
50	45	9	10	1	1
54	81	6	1	5	25
80	60	1	6	5	25
75	68	2.5	3.5	1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				ΣD^2	72

$$R=1-\frac{6\left[72+\frac{1}{12}(2^3-2)+\frac{1}{12}(3^3-3)+\frac{1}{12}(2^3-2)\right]}{N^3-N}$$

$$= 1-\frac{6\left[72+\frac{1}{12}+2+\frac{1}{12}\right]}{10^3-10}$$

$$=1-\frac{6\times[72+3]}{990}$$

$$=1-\frac{6\times75}{990}$$

$$= 1-\frac{450}{990} = 1 - 0.4545$$

$$= \underline{\underline{0.5455}}$$

Merits of Rank Correlation method

1. Rank correlation coefficient is only an approximate measure as the actual values are not used for calculations

2. It is very simple to understand the method.
3. It can be applied to any type of data, ie quantitative and qualitative
4. It is the only way of studying correlation between qualitative data such as honesty, beauty etc.
5. As the sum of rank differences of the two qualitative data is always equal to zero, this method facilitates a cross check on the calculation.

Demerits of Rank Correlation method

1. Rank correlation coefficient is only an approximate measure as the actual values are not used for calculations.
2. It is not convenient when number of pairs (ie. N) is large
3. Further algebraic treatment is not possible.
4. Combined correlation coefficient of different series cannot be obtained as in the case of mean and standard deviation. In case of mean and standard deviation, it is possible to compute combine arithematic mean and combined standard deviation.

Concurrent Deviation Method:

Concurrent deviation method is a very simple method of measuring correlation. Under this method, we consider only the directions of deviations. The magnitudes of the values are completely ignored. Therefore, this method is useful when we are interested in studying correlation between two variables in a casual manner and not interested in degree (or precision).

Under this method, the nature of correlation is known from the direction of deviation in the values of variables. If deviations of 2 variables are concurrent, then they move in the same direction, otherwise in the opposite direction.

The formula for computing the coefficient of concurrent deviation is: -

$$r = \pm\sqrt{\pm\frac{(2c-N)}{N}}$$

Where N = No. of pairs of symbol

C = No. of concurrent deviations (ie, No. of + signs in 'dx dy' column)

Steps:

1. Every value of 'X' series is compared with its proceeding value. Increase is shown by '+' symbol and decrease is shown by '-'
2. The above step is repeated for 'Y' series and we get 'dy'
3. Multiply 'dx' by 'dy' and the product is shown in the next column. The column heading is 'dxdy'.

4. Take the total number of '+' signs in 'dxdy' column. '+' signs in 'dxdy' column denotes the concurrent deviations, and it is indicated by 'C'.

5. Apply the formula:

$$r = \pm\sqrt{\pm\left(\frac{2c-N}{N}\right)}$$

If 2c> $N, then\ r = +ve\ and\ if\ 2c < N, then\ r = -ve.$

Qn:- Calculate coefficient if correlation by concurrent deviation method:-

Year	:	2003	2004	2005	2006	2007	2008	2009	2010	2011
Supply	:	160	164	172	182	166	170	178	192	186
Price	:	292	280	260	234	266	254	230	190	200

Sol: Computation of coefficient of concurrent

	Deviation			
Supply (x)	Price (y)	dx	dy	dxdy
160	292	+	-	-
164	280	+	-	-
172	260	+	-	-
182	234	+	-	-
166	266	-	+	-
170	254	+	-	-
178	230	+	-	-
192	190	+	-	-
186	200	-	+	-
				C = 0

$$\mathbf{r} = \pm\sqrt{\pm\frac{(2C-N)}{N}}$$

$$= \pm\sqrt{\pm\frac{(2\times0)-8}{8}}$$

$$= \pm\sqrt{\frac{0-8}{8}} = = \pm\sqrt{\frac{-8}{8}} = \underline{-1}$$

Merits of concurrent deviation method:

1. It is very easy to calculate coefficient of correlation
2. It is very simple understand the method
3. When the number of items is very large, this method may be used to form quick idea about the degree of relationship
4. This method is more suitable, when we want to know the type of correlation (ie, whether positive or negative).

Demerits of concurrent deviation method:

1. This method ignores the magnitude of changes. Ie. Equal weight is give for small and big changes.
2. The result obtained by this method is only a rough indicator of the presence or absence of correlation
3. Further algebraic treatment is not possible
4. Combined coefficient of concurrent deviation of different series cannot be found as in the case of arithmetic mean and standard deviation.

CHAPTER - 3

REGRESSION ANALYSIS

Introduction:-

Correlation analysis analyses whether two variables are correlated or not. After having established the fact that two variables are closely related, we may be interested in estimating the value of one variable, given the value of another. Hence, regression analysis means to analyse the average relationship between two variables and thereby provides a mechanism for estimation or predication or forecasting.

The term 'Regression" was firstly used by Sir Francis Galton in 1877. The dictionary meaning of the term 'regression" is "stepping back" to the average.

Definition:

"Regression is the measure of the average relationship between two or more variables in terms of the original units of the date".

"Regression analysis is an attempt to establish the nature of the relationship between variables-that is to study the functional relationship between the variables and thereby provides a mechanism for prediction or forecasting".

It is clear from the above definitions that Regression Analysis is a statistical device with the help of which we are able to estimate the unknown values of one variable from known values of another variable. The variable which is used to predict the another variable is called independent variable (explanatory variable) and, the variable we are trying to predict is called dependent variable (explained variable).

The dependent variable is denoted by X and the independent variable is denoted by Y.

The analysis used in regression is called simple linear regression analysis. It is called simple because three is only one predictor (independent variable). It is called linear because, it is assumed that there is linear relationship between independent variable and dependent variable.

Types of Regression:-

There are two types of regression. They are linear regression and multiple regression.

Linear Regression:

It is a type of regression which uses one independent variable to explain and/or predict the dependent variable.

Multiple Regression:

It is a type of regression which uses two or more independent variable to explain and/or predict the dependent variable.

Regression Lines:

Regression line is a graphic technique to show the functional relationship between the two variables X and Y. It is a line which shows the average relationship between two variables X and Y.

If there is perfect positive correlation between 2 variables, then the two regression lines are winding each other and to give one line. There would be two regression lines when there is no perfect correlation between two variables. The nearer the two regression lines to each other, the higher is the degree of correlation and the farther the regression lines from each other, the lesser is the degree of correlation.

Properties of Regression lines:-

1. The two regression lines cut each other at the point of average of X and average of Y (i.e $\overline{X}$ and $\overline{Y}$)
2. When r = 1, the two regression lines coincide each other and give one line.
3. When r = 0, the two regression lines are mutually perpendicular.

Regression Equations (Estimating Equations)

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, therefore two regression equations. They are :-

1. Regression Equation of X on Y:- This is used to describe the variations in the values of X for given changes in Y.
2. Regression Equation of Y on X :- This is used to describe the variations in the value of Y for given changes in X.

Least Square Method of computing Regression Equation:

The method of least square is an objective method of determining the best relationship between the two variables constituting a bivariate data. To find out best relationship means to determine the values of the constants involved in the functional relationship between the two variables. This can be done by the principle of least squares:

The principle of least squares says that the sum of the squares of the deviations between the observed values and estimated values should be the least. In other words, $\Sigma(y - y_C)^2$ will be the minimum.

With a little algebra and differential calculators we can develop some equations (2 equations in case of a linear relationship) called normal equations. By solving these normal equations, we can find out the best values of the constants.

Regression Equation of Y on X:-

$$Y = a + bx$$

The normal equations to compute 'a' and 'b' are: -

$$\Sigma y = Na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Regression Equation of X on Y:-

$$X = a + by$$

The normal equations to compute 'a' and 'b' are:-

$$\Sigma x = Na + n\Sigma y$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

Qn:- Find regression equations x and y and y on x from the following:-

X:	25	30	35	40	45	50	55
Y:	18	24	30	36	42	48	54

Sol: Regression equation x on y is:

x = a + by

Normal equations are:

$$\Sigma x = Na + b\Sigma y$$

$$\Sigma xy = a\Sigma y + b\Sigma y2$$

Computation of Regression Equations				
x	**y**	**x^2**	**y^2**	**xy**
25	18	625	324	450
30	24	900	576	720
35	30	1225	900	1050
40	36	1600	1296	1440
45	42	2025	1764	1890
50	48	2500	2304	2400
55	54	3025	2916	2970
$\Sigma x = 280$	$\Sigma y = 252$	$\Sigma x2 = 11900$	$\Sigma y2 = 10080$	$\Sigma xy = 10920$

280 = 7a+ 252 b ----------(1)

10920 = 252a+10080 b ----------(2)

Eq. 1x 36 $\Longrightarrow$ 10080 = 252a + 9072b ----------(3)

10920 = 252a + 10080b ---------- (2)

(2) x (3) $\Longrightarrow$ 840 = 0 + 1008 b

1008 b = 840

b = $\frac{840}{1008}$ = 0.83

Substitute b = 0.83 in equation (1)

280 = 7a + (252×0.83)

280 = 7 a + 209.16

7a+ 209.116 = 280

7a = 280-209.160

a = $\frac{70.84}{7}$ = 10.12

Substitute a = 10.12 and b =0.83 in regression equation:

$\underline{X = 10.12 + 0.83\ y}$

Regression equation Y on X is:

y = a + bx

Normal Equations are:-

$\Sigma y = Na + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

252 = 7a + 280 b ------- (1)

10920 = 280 a+ 11900 b ------- (2)

(1) x 40 ⟹ 10080 = 280 a + 11200 b ------- (3)

10920 = 280 a+ 11900 b ------- (2)

(2) – (3) ⟹ 840 = 0 + 700 b

700 b = 840

b = $\frac{840}{700}$ = 1.2

Substitute b = 1.2 in equation (1)

252 = 7a + (280x1.2)

252 = 7a + 336

7a + 336 = 252

$7a = 252 - 336 = -84$

$a = \frac{-84}{7} = -12$

Substitute a = -12 and b = 1.2 in regression equation

y = -12+1.2x

Qn:- From the following bivariate data, you are required to: -

(a) Fit the regression line Y on X and predict Y if x = 20

(b) Fir the regression line X on Y and predict X if y = 10

X:	4	12	8	6	4	4	16	8
Y:	14	4	2	2	4	6	4	12

Computation of regression equations				
x	**y**	**x2**	**y2**	**xy**
4	14	16	196	56
12	4	144	16	48
8	2	64	4	16
6	2	36	4	12
4	4	16	16	16
4	6	16	36	24
16	4	256	16	64
8	12	64	144	96
$\Sigma x = 62$	$\Sigma y = 48$	$\Sigma x^2 = 612$	$\Sigma y^2 = 432$	$\Sigma xy = 332$

Regression equation y on x

y = a + bx

Normal equations are:

$$\Sigma y = Na + b\Sigma y$$

$$\Sigma xy = a\Sigma \times b\Sigma x^2$$

$48 = 8a + 62\ b$ ……(1)

$332 = 62a + 612\ b$ ……(2)

eq. (1) $\times 62 \implies 2{,}976 = 496a + 3844\ b$ ….. (3)

eq. (2) $\times 8 \implies 2{,}976 = 496a + 4896\ b$ ….. (4)

eq. (3) $\times$ eq. (4) $\implies 320 = 0 + -1052b$

$-1052\ b = 320$

$b = \frac{320}{-1052}$

Substitute b = -0.304 in eq (1)

$48 = 8a + (62 \times -0.304)$

$48 = 8a + -18.86$

$48 + 18.86 = 8a$

$a = 66.86$

$a = \frac{66.86}{8} = \underline{8.36}$

Substitute a = 8.36 and b = -0.304 in regression equation y on x :

$y = 8.36 + -0.3042\ x$

$y = 8.36 - 0.3042\ x$

If x = 20, then,

$y = 8.36 - (0.3042 \times 20)$

$= 8.36 - 6.084$

$= \underline{2.276}$

(b) Regression equation X on Y:

$X = a + by$

Normal equations:

$$\Sigma x = Na + b\Sigma y$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

62 = 8a + 48 b …….(1)

332 = 48 a + 432 b …….(2)

eq (1) × 6 $\Longrightarrow 372 = 48a + 288b$ ….(3)

332 = 48 a + 432 b ……(2)

eq (2) − (3) $\Longrightarrow -40 = 0 + 144b$

144 b = -40

b = $\frac{-40}{144}$ = - 0.2778

Substitute b = -0.2778 in equation (1)

62 = 8a + (48× $\bar{0}.2778$)

62 = 8a + -13.3344

62+13.3344 = 8 a

8a = 75.3344

a = $\frac{75.3344}{8}$ = 9.4168

Substitute a = 9.4168 and b = -0.2778 in regression equation:

x = 9.4168 + -0.2778 y

x = 9.4168 + -0.2778 y

If y=10, then

x=9.4168 – (0.2778x10)

x= 9.4168 – 2.778

x = 6.6388

Regression Coefficient method of computing Regression Equations:

Regression equations can also be computed by the use of regression coefficients. Regression coefficient X on Y is denoted as b_{xy} and that of Y on X is denoted as b_{yx}.

Regression Equation x on y:

$x - \bar{x} = b_{xy}(y - \bar{y})$

i.e $x - \bar{x} = r.\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

Regression Equation y on x:

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

i.e $y - \bar{y} = r.\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

Properties of Regression Coefficient:

1. There are two regression coefficients. They are b_{xy} and b_{yx}
2. Both the regression coefficients must have the same signs. If one is +ve, the other will also be a +ve value.
3. The geometric mean of regression coefficients will be the coefficient of correlation. $r = \sqrt{b_{xy}.b_{yx}}$
4. If x $\tilde{x}$ and $\tilde{y}$ are the same, then the regression coefficient and correlation coefficient will be the same.

Computation of Regression Co-efficients

Regression co-efficients can be calculated in 3 different ways:

1. Actual mean method
2. Assumed mean method
3. Direct method

Actual mean method:-

Regression coefficient x on y $(b_{xy}) = \frac{\sum xy}{\sum y^2}$

Regression coefficient y on x $(b_{yx}) = \frac{\sum xy}{\sum x^2}$

Where $x = x - \bar{x}$

$y = y - \bar{y}$

Assumed mean method:

Regression coefficient x on y (b_{xy}) } $\frac{\Sigma dxdy - (\Sigma dx).(\Sigma dy)}{\Sigma dy^2 - (\Sigma dy)^2}$

Regression coefficient y on x (b_{yx}) } $\frac{\Sigma dxdy - (\Sigma dx).(\Sigma dy)}{\Sigma dy^2 - (\Sigma dy)^2}$

Where dx = deviation from assumed mean of X

dy = deviation from assumed mean of Y

Direct method:-

Regression Coefficient x on y (b_{xy}) $\left.\right\}$ $\frac{N\Sigma xy-\Sigma x.\Sigma y}{N\Sigma y^2-(\Sigma x)^2}$

Regression Coefficient y on x (b_{yx}) $\left.\right\}$ $\frac{N\Sigma xy-\Sigma x.\Sigma y}{N\Sigma x^2-(\Sigma x)^2}$

Qn:- Following information is obtained from the records of a business organization:-

Sales (in '000):	91	53	45	76	89	95	80	65
Advertisement Expense (₹in '000)	15	8	7	12	17	25	20	13

You are required to:-

1. Compute regression coefficients under 3 methods
2. Obtain the two regression equations and
3. Estimate the advertisement expenditure for a sale of Rs. 1,20,000

Let x = sales

y = Advertisement expenditure

Computation of regression Coefficients under actual mean method

x	y	x - $\bar{x}$	y - $\bar{y}$	xy	x^2	y^2
91	15	16.75	0.375	6.28	280.56	0.14
53	8	-21.65	-6.625	140.78	451.56	43.89
45	7	-29.25	-7.625	223.03	855.56	58.14
76	12	1.75	-2,625	-4.59	3.06	6.89
89	17	14.75	-2.375	35.03	217.56	5.64
95	25	20.75	10.375	215.28	430.56	107.64
80	20	5.75	5.375	30.91	33.06	28.89
65	13	-9.25	-1.625	15.03	85.56	2.64
$\Sigma x = 594$	$\Sigma y = 117$			$\Sigma xy = 661.75$	$\Sigma x2 = 2357.48$	$\Sigma y = 253.87$

$\overline{X} = \frac{\Sigma X}{N} = \frac{294}{8} = 74.25$

$\overline{Y} = \frac{\Sigma Y}{N} = \frac{117}{8} = 14.625$

Regression coefficient x on y (b_{xy}) $\Big\}$ $\frac{\Sigma xy}{\Sigma y^2}$

$= \frac{661.75}{253.87} = 2.61$

Regression coefficient Y on X $((b_{yx})$ $\Big\}$ $\frac{\Sigma xy}{\Sigma x^2}$

$= \frac{661.75}{2357.48} = 0.28$

Computation of Regression Coefficient under assured mean method

x	y	x-70 (dx)	y-15 (dy)	dxdy	dx^2	dy^2
91	15	21	0	0	441	0
53	8	-17	-7	+119	289	49
45	7	-25	-8	+200	625	64
76	12	6	-3	-18	36	9
89	17	19	2	+38	361	4
95	25	25	10	+250	625	100
80	20	10	5	+50	100	25
65	13	-5	-2	+10	25	4
		$\Sigma dx = 34$	$\Sigma dy = -3$	$\Sigma dxdy = 649$	$\Sigma dx^2 = 2502$	$\Sigma dx^2 = 255$

Regression Coefficient x on y (b_{xy}) $\Big\}$ $\frac{N\Sigma dxdy - \Sigma dx.\Sigma dy}{N\Sigma dy^2 - (\Sigma dy)^2}$

$$= \frac{8 \times 649) - (34 \times -3)}{(8 \times 255) - (-3)^2}$$

$$= \frac{5192 - -102}{2040 - 9}$$

$$= \frac{5192+102}{2031} = \frac{5294}{2031}$$

$$= 2.61$$

Regression coefficient y on x (b_{yx})

$$\frac{N\Sigma dxdy - \Sigma dx.\Sigma dy}{N\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{(8 \times 649) - (34 \times -3)}{(8 \times 2502) - (34)^2}$$

$$= \frac{5192 - -102}{20016 - 1156}$$

$$= \frac{5294}{18860}$$

$$= 0.28$$

Computation of Regression Coefficient under direct method				
x	**y**	**xy**	**x^2**	**y^2**
91	15	1365	8281	225
53	8	424	2809	64
45	7	315	2025	49
76	12	912	5776	144
89	17	1513	7921	289
95	25	2375	9025	625
80	20	1600	6400	400
65	13	845	4225	169
$\Sigma x = 594$	$\Sigma y = 117$	$\Sigma xy = 9349594$	$\Sigma x^2 = 46462$	$\Sigma y^2 = 1965$

Regression coefficient x on y (b_{xy}) $\left.\right\}$ $\frac{N\Sigma xy-\Sigma x.\Sigma y}{N\Sigma y)^2-(\Sigma y)^2}$

$= \frac{(8\times9349)-(594\times117)}{(8\times1965)-117^2}$

$= \frac{74792-69498}{15720-13689}$ $\quad \frac{5294}{2031}$

$= \underline{2.61}$

Regression coefficient y on x (b_{yx}) $\left.\right\}$ $=\frac{N\Sigma xy-\Sigma x.\Sigma y}{N\Sigma x^2-(\Sigma x)^2}$

$\frac{(8\times9349)-(594\times117)}{(8\times46462)-594)^2}$

$\frac{74792-69498}{371696-352836}$

$\frac{5294}{18860} = 0.28$

3) a) Regression equation X on Y:

$(x-\bar{x}) = b_{xy}\ (y-\bar{y})$

$(x-74.25) = 2.61\ (\bar{y}-14.625)$

$(x-74.25) = 2.61\ y-38.17$

$x = 74.25 - 38.17+2.61y$

$x = \underline{36.08 + 2.61y}$

b) Regression equation y pm x:

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

$(y-14.625) = 0.28\ (x-74.25)$

$y = 14.625 = 0.28\times -20.79$

$y = 14.625 - 20.79 + 0.28\times$

$y = -6.165+0.28\times$

$\underline{y = 0.28\times -6.165}$

4) If sales (x)is Rs. 1,20,000, then

Estimated advertisement Exp (y) = (0.28x120)-6.165

= (33.6 – 6.165

= 27.435

i.e Rs. 27,435

Qn: In a correlation study, the following values are obtained:

	x	y
Mean	65	67
Standard deviation	2.5	3.5
Coefficient of correlation	0.8	

Find the regression equations

Sol: Regression equation x on y is:

$x-\bar{x} = b_{xy}\,(y-\bar{y})$

$x-\bar{x} = r \cdot \frac{\sigma_x}{\sigma_y}\,(y-\bar{y})$

$x - 65 = 0.8 \times \frac{2.5}{3.5}\,(y-67)$

$x - 65 = 0.5714\,(y-67)$

$x - 65 = 0.5714y-38.2838$

$x = 65 – 38.2838+0.5714y$

$x = 26.72 + 0.5714y$

Regression equation y on x is:

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

$y - \bar{y} = r \cdot \frac{\sigma_y}{\sigma_x}\,(x-\bar{x})$

$y - 67 = 0.8 \times \frac{3.5}{2.5}\,(x-65)$

$y - 67 = 1.12\,(x-65)$

$y = 67 – (1.12 \times 65) = 1.12\,x$

$y = 67.72.8 + 1.12x$

y = -5.8 +1.12x

y = 1.12x - 5.8

Qn: Two variables gave the following data

$\bar{x}$ = 20, σ_x = 4, r = 0.7

$\bar{y}$ = 15, σ_y = 3

Obtain regression lines and find the most likely value of y when x=24

Sol: Regression Equation x on y is

$(x - \bar{x}) = b_{xy} (y - \bar{y})$

$(x - \bar{x}) = r \cdot \frac{\sigma_x}{\sigma_y} (y - \bar{y})$

$(x - 20) = 0.7 \times \frac{4}{3} (y-15)$

$(x - 20) = \frac{2.8}{3} (y-15)$

(x - 20) = 0.93(y-15)

x = 20 + 0.93y – 13.95

x = 20 - 13.95 + 0.93y

x = 6.05 + 0.93y

Regression Equation y on x is

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

$(y - \bar{y}) = r \cdot \frac{\sigma_y}{\sigma_x} (x - \bar{x})$

$(y - \bar{y}) = 0.7 \times \frac{3}{4} (x - 20)$

(y -15) = 0.525(x – 20)

y – 15 = 0.525 x -10.5

y = 15-10.5+ 0.525x

y =4.5+0.525x

If X = 24, then

y = 4.5 + (0.525×24)

y = 4.5 + 12.6

y = 17.1

Qn: For a given set of bivariate data, the following results were obtained:

$\bar{x}$ = 53.2, $\bar{y}$ = 27.9, b_{yx} = -1.5 and b_{xy} = -0.2

Find the most probable value of y when x = 60. Also find 'r'.

Sol: Regression Equation y on x is:

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

(y - 27.9) = -1.5 (x-53.2)

(y - 27.9) = -1.5x + 79.8

y = 79.8+27.9 – 1.5×

y = 107.7 – 1.5×

If x = 60, then

y = 107.7 – (1.5×60)

= 107.7-90

= 17.7

r $= \sqrt{b_{xy} \times b_{yx}}$

$= -\sqrt{1.5 \times 0.2} = -\sqrt{30}$ = -0.5477

Correlation		Regression
1	It studies degree of relationship between variables	It studies the nature of relationship between variables
2	It is not used for prediction purposes	It is basically used for prediction purposes
3	It is basically used as a tool for determining the degree of relationship	It is basically used as a tool for studying cause and effect relationship
4	There may be nonsense correlation between two variables	There is no such nonsense regression
5	There is no question of dependent and independent variables	There must be dependent and independent variables

CHAPTER - 4

THEORY OF PROBABILITY

INTRODUCTION

Probability refers to the chance of happening or not happening of an event. In our day today conversations, we may make statements like " probably he may get the selection", "possibly the Chief Minister may attend the function", etc. Both the statements contain an element of uncertainly about the happening of the even. Any problem which contains uncertainty about the happening of the event is the problem of probability.

Definition of Probability

The probability of given event may be defined as the numerical value given to the likely hood of the occurrence of that event. It is a number lying between '0' and '1' '0' denotes the even which cannot occur, and '1' denotes the event which is certain to occur. For example, when we toss on a coin, we can enumerate all the possible outcomes (head and tail), but we cannot say which one will happen. Hence, the probability of getting a head is neither 0 nor 1 but between 0 and 1. It is 50% or ½

Terms use in Probability.

Random Experiment

A random experiment is an experiment that has two or more outcomes which vary in an unpredictable manner from trial to trail when conducted under uniform conditions.

In a random experiment, all the possible outcomes are known in advance but none of the outcomes can be predicted with certainty. For example, tossing of a coin is a random experiment because it has two outcomes (head and tail), but we cannot predict any of them which certainty.

Sample Point

Every indecomposable outcome of a random experiment is called a sample point. It is also called simple event or elementary outcome.

Eg. When a die is thrown, getting '3' is a sample point.

Sample space

Sample space of a random experiment is the set containing all the sample points of that random experiment.

Eg:- When a coin is tossed, the sample space is (Head, Tail)

Event

An event is the result of a random experiment. It is a subset of the sample space of a random experiment.

Sure Event (Certain Event)

An event whose occurrence is inevitable is called sure even.

Eg:- Getting a white ball from a box containing all while balls.

Impossible Events

An event whose occurrence is impossible, is called impossible event. Eg:- Getting a white ball from a box containing all red balls.

Uncertain Events

An event whose occurrence is neither sure nor impossible is called uncertain event.

Eg:- Getting a white ball from a box containing white balls and black balls.

Equally likely Events

Two events are said to be equally likely if anyone of them cannot be expected to occur in preference to other. For example, getting herd and getting tail when a coin is tossed are equally likely events.

Mutually exclusive events

A set of events are said to be mutually exclusive of the occurrence of one of them excludes the possibiligy of the occurrence of the others.

Exhaustive Events:

A group of events is said to be exhaustive when it includes all possible outcomes of the random experiment under consideration.

Dependent Events:

Two or more events are said to be dependent if the happening of one of them affects the happening of the other.

PERMUTATIONS

Permutation means arrangement of objects in a definite order. The number of arrangements (permutations) depends upon the total number of objects and the number of objects taken at a time for arrangement.

The number of permutations is calculated by using the following formula:

$$^{n}P_{r=} \quad \frac{n!}{(n-r)!}$$

$!=$ Factorial

$n=$ Total number of objects

$r=$ Number of objects taken at a time for arrangement

If whole the objects are taken at a time for arrangement, then number of permutations is calculated by using the formula :

$${}^{n}P_{r} = {}^{n}P_{n}$$

$$= \frac{n!}{(n-n)!} \quad \frac{n!}{o!} = \frac{n!}{1} = n!$$

$$\boxed{{}^{n}P_{n} = n!}$$

Question:-

A factory manager purchased 3 new machines, A, B and C. How many number of times he can arrange the 3 machines?

Solution :

$${}^{n}P_{r=} \quad \frac{n!}{(n-r)!}$$

$n =$ 3

$r =$ 3

Here 'n' and 'r' are same

$\therefore \quad {}^{n}P_{r=} n!$

$= 3! = 3 \times 2 \times 1 = 6.$

Question :

In how many ways 3 people be seated on a bench if only two seats are available.

Solution

$${}^{n}P_{r=} \quad \frac{n!}{(n-r)!}$$

$n =$ 3

$r =$ 2

$\therefore \quad 3P_2 = \frac{3!}{(3-2)!} = \frac{3!}{1!} = \frac{3\times2\times1}{1} = 6$ ways

Computation of Permutation when objects are alike:-

Sometimes, some of the objects of a group are alike. In such a situation number of permutations is calculated as :-

$${}^{n}P_{r=} \quad \frac{n!}{n,!\, n_2!\, n_3! \ldots\ldots n_n!}$$

$n_1 =$ number of alike objects in first category

$n_2 =$ number of alike objects in second category

If all items are alike, you know, they can be arranged in only one order

ie , ${}^nP_r = \frac{n!}{n_1!}$

Question:

Find the number of permutations of letters in the word 'COMMUNICATION'

Solution

$${}^nP_r = \frac{n!}{n,!\, n_2!\, n_3! \ldots \ldots .}$$

C = 2, 0 = 2; M = 2 ; U = 1; N = 2;

1 = 2; A =1; T = 1

$$\therefore \quad {}^nP_r = \frac{13!}{2!2\ !2!1\ !2\ !2\ !1!1\ !}$$

$$= \frac{13 \times 12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1 \times 2 \times 1 \times 1 \times 2 \times 1 \times 2 \times 1 \times 1 \times 1}$$

= 194594400 times

COMBINATIONS

Combination means selection or grouping of objects without considering their order. The number of combinations is calculated by using the following formula:

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

Question

A basket contains 10 mangoes. In how many ways 4 mangoes from the basket can be selected?

Solution

$${}^nC_r = \frac{n!}{(n-r)!\, r!}$$

n = 10

r = 4

$${}^{10}C_4 = \frac{10!}{(10-4)!4!} = \frac{10!}{6!4!} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1}$$

= 210 ways

Questions

How many different sets of 5 students can be chosen out of 20 qualified students to represent a school in an essay context ?

Solution

$$^{n}C_{r} = \frac{n!}{(n-r)!\,r!}$$

n = 20

r = 5

$$^{20}C_{5} = \frac{20!}{(20-5)!5!} = \frac{20!}{15!5!} = \frac{20 \times 19 \times 18 \times 17 \times 16}{4 \times 3 \times 2 \times 1}$$

= 15504 Sets

DIFFERENT SCHOOLS OF THOUGHT ON PROBABILITY

Different Approaches/Definitions of Probability

There are 4 important schools of thought on probability :-

1. Classical or Priori Approach } Objective Probability Approach
2. Relative frequency or Empirical Approach
3. Subjective or Personalistic Approach
4. Modern or Axiomatic Approach

1. Classical or Priori Approach

If out of '*n*' exhaustive, mutually exclusive and equally likely outcomes of an experiment; '*m*' are favourable to the occurrence of an event 'A', then the probability of 'A' is defined as to be $\frac{m}{n}$.

$$P(A) = \frac{m}{n}$$

According to Laplace, a French Mathematician, " the probability is the ratios of the number of favourable cases to the total number of equally likely cases."

$$P(A) = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ equally\ likely\ cases}$$

Question

What is the chance of getting a head when a coin is tossed?

Solution

Total number of cases = 2

No. of favorable cases = 1

Probability of getting head = $\frac{1}{2}$

Question

A die is thrown. Find the probability of getting.

(1) A '4'

(2) an even number

(3) '3' or '5'

(4) less than '3'

Solution

Sample space is (1,2, 3, 4, 5, 6)

(1) Probability (getting '4) = $\frac{1}{6}$

(2) Probability (getting an even number) = $\frac{3}{6} = \frac{1}{6}$

(3) Probability (getting 3 or 5) = $\frac{2}{6} = \frac{1}{3}$

(4) Probability (getting less than '3') = $\frac{2}{6} = \frac{1}{3}$

Question

A ball is drawn from a bag containing 4 white, 6 black and 5 yellow balls. Find the probability that a ball drawn is :-

(1) White (2) Yellow (3) Black (4) Not yellow (5) Yellow or white

Solution

(1)P (drawing a white ball) = $\frac{4}{15}$

(2) P (drawing a yellow ball) = $\frac{5}{15} = \frac{1}{3}$

(3) P (drawing a black ball) = $\frac{6}{15} = \frac{2}{5}$

(4) P (drawing not a yellow ball) = $\frac{10}{15} = \frac{2}{3}$

(5) P (drawing a yellow or white ball) = $\frac{9}{15} = \frac{3}{5}$

Question

There are 19 cards numbered 1 to 19 in a box. If a person drawn one at random, what is the probability that the number printed on the card be an even number greater than 10?

Solution

The even numbers greater than 10 are 12, 14, 16 and 18.

$\therefore$P (drawing a card with an even number greater than 10) $= \frac{4}{9}$

Question

Two unbiased dice are thrown. Find the probability that :-

(a) Both the dice show the same number

(b) One die shows 6

(c) First die shows 3

(d) Total of the numbers on the dice is 9

(e) Total of the numbers on the dice is greater than 8

(f) A sum of 11

Solution

When 2 dice are thrown the sample space consists of the following outcomes :-

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

(a) P(that both the dice shows the same number) $= \frac{6}{36} = \frac{1}{6}$

(b) P (that one die shows 6) $= \frac{10}{36} = \frac{5}{18}$

(c) P (that first die shows 3) $= \frac{6}{36} = \frac{1}{6}$

(d) P (that total of the numbers on the dice is 9) $= \frac{4}{36} = \frac{1}{9}$

(e) P (that total of the number is greater than 8) $= \frac{10}{36} = \frac{5}{18}$

(f) P (that a sum of 11) $= \frac{2}{36} = \frac{1}{18}$

Problems based on combination results

Question

A box contains 6 white balls and 4 green balls. What is the probability of drawing a green ball?

Solution

Probable number of cases = ${}^4C_1$

Total number of cases = ${}^{10}C_1$

P(drawing a green ball) $= \frac{{}^4C_1}{{}^{10}C_1}$

$$= \frac{\frac{4!}{(4-1)! \times 1!}}{\frac{10\,!}{(10-1)! \times 1!}}$$

$$= \frac{\frac{4!}{3! \times 1!}}{\frac{10!}{9! \times 1!}}$$

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

$$= \frac{4}{10} = \frac{2}{5}$$

Question

What is the probability of getting 3 red balls in a draw of 36 balls from a bag containg 5 red balls and 46 black balls?

Solution

Favourable number of cases = ${}^5C_3$

Total number of cases = ${}^9C_3$

P(getting 3 white balls)= $\frac{{}^5C_3}{{}^9C_3}$

$$= \frac{\frac{5!}{(5-3)!\,3!}}{\frac{9!}{(9-3)!\,3!}} = \frac{\frac{5!}{2!\,3!}}{\frac{9!}{6!\,3!}}$$

$$= \frac{\frac{5 \times 4}{2}}{\frac{9 \times 8 \times 7}{3 \times 2 \times 1}} = \frac{\frac{20}{2}}{3 \times 4 \times 7} = \frac{10}{84}$$

$$= \frac{5}{42}$$

Question

A committee is to be constituted by selecting three people at random from a group consisting of 5 Economists and 4 Statisticians. Find the probability that the committee will consist of :

(a)3 Economists

(b) 3 Statisticians

(c)3 Economists and 1 Statistician

(d) 1 Economist and 2 Statistician

(a) P (Selecting 3 Economists) = $\frac{{}^5C_3}{{}^9C_3}$

$$= \frac{\frac{5!}{(5-3)!3!}}{\frac{9!}{(9-3)!3!}} = \frac{\frac{5!}{2!\ 3!}}{\frac{9!}{6!\ 3!}}$$

$$\frac{\frac{5\times4}{2\times1}}{\frac{9\times8\times7}{3\times2\times1}} = \frac{10}{84} = \frac{5}{42}$$

(b) P(Selecting 3 Statisticians) = $\frac{{}^4C_3}{{}^9C_3}$

$$\frac{\frac{4!}{(4-3)!3!}}{\frac{9!}{(9-3)!3!}} = \frac{\frac{4!}{1!\ 3!}}{\frac{9!}{6!\ 3!}}$$

$$\frac{4}{84} = \frac{1}{21}$$

(c) P (Selecting 2 Economists and 1 Statistician) $\Big\}$ $= \frac{{}^5C_2 \times {}^4C_1}{{}^9C_3}$

$$= \frac{\frac{5!}{(5-2)!2!}\times\frac{4!}{(4-1)!1!}}{\frac{9!}{(9-3)!\,3!}}$$

$$= \frac{\frac{5!}{3!\ 2!}\times\frac{4!}{3!\ 1!}}{\frac{9!}{6!\ 3!}}$$

$$= \frac{\frac{5\times4}{2\times1}\times\frac{4}{1}}{\frac{9\times8\times7}{3\times2\times1}} = \frac{10\times4}{84}$$

$$= \frac{40}{84} = \frac{10}{21}$$

(d) P(Selecting 1 Economist and 2 Statistician $\Big\}$ $= \frac{{}^5C_1 \times {}^4C_2}{{}^9C_3}$

$$= \frac{\frac{5!}{(5-1)!\,1!} \times \frac{4!}{(4-2)!2!}}{\frac{9!}{(9-3)!\,3!}}$$

$$= \frac{5 \times 6}{84} = \frac{30}{84} = \frac{5}{14}$$

Questions

A committee of 5 is to be formed from a group of 8 boys and 7 girls. Find the probability that the committee consists of at least one girl.

Solution

P (that committee consists of at least one girl) = P(one girl & 4 boys) or P(2 girls & 3 boys) or P(3 girls & 2 boys) or P(4girls & 1 boy) or P(5 girls)

$$= \frac{({}^7C_1 \times {}^8C_4) + ({}^7C_2 \times {}^8C_3) + ({}^7C_3 \times {}^8C_2) + ({}^7C_4 \times {}^8C_1) + {}^7C_5}{{}^{15}C_5}$$

$$= \frac{\left(\frac{7!}{6!\,1!} \times \frac{8!}{4!\,4!}\right) + \left(\frac{7!}{5!\,2!} \times \frac{8!}{5!\,3!}\right) + \left(\frac{7!}{4!\,3!} \times \frac{8!}{6!\,2!}\right) + \left(\frac{7!}{3!\,1!} \times \frac{8!}{7!1!}\right) + \frac{7!}{2!\,5!}}{\frac{15!}{10!5!}}$$

$$= \frac{\left(7 \times \frac{8 \times 7 \times 6 \times 5}{4 \times 3 \times 2}\right) + \left(\frac{7 \times 6}{2} \times \frac{8 \times 7 \times 6}{3 \times 2}\right) + \left(\frac{7 \times 6 \times 5!}{3 \times 2} \times \frac{8 \times 7}{2}\right) + \left(\frac{7 \times 6 \times 5}{3 \times 2} \times 8\right) + \frac{7 \times 6}{2}}{\frac{15 \times 14 \times 13 \times 12 \times 11}{5 \times 4 \times 3 \times 2}}$$

$$= \frac{(7 \times 70) + (21 \times 56) + (35 \times 28) + (35 \times 8) + 21}{3003}$$

$$= \frac{490 + 1176 + 980 + 280 + 21}{3003} = \frac{2947}{3003}$$

This problem can be solved in the following method also.

P(that the committee consists of at least one girl) = 1 – P (that the committee consists of all boys

$$= 1 - \left(\frac{{}^8C_5}{{}^{15}C_5}\right)$$

$$= 1 - \frac{\frac{8!}{3!\,5!}}{\frac{15!}{10!\,5!}} = 1 - \frac{\frac{8 \times 7 \times 6}{3 \times 2}}{\frac{15 \times 14 \times 13 \times 12 \times 11}{5 \times 4 \times 3 \times 2}}$$

$$= 1 - \frac{56}{3003}$$

$$= \frac{3003 - 56}{3003} = \frac{2947}{3003}$$

Limitations of Classical Definition:

1. Classical definition has only limited application in coin-tossing die throwing etc. It fails to answer question like "What is the probability that a female will die before the age of 64?"
2. Classical definition cannot be applied when the possible outcomes are not equally likely. How can we apply classical definition to find the probability of rains? Here, two possibilities are "rain" or "no rain". But at any given time these two possibilities are not equally likely.
3. Classical definition does not consider the outcomes of actual experimentations.

Relative Frequency Definition or Empirical Approach

According to Relative Frequency definition, the probability of an event can be defined as the relative frequency with which it occurs in an indefinitely large number of trials.

If an even 'A' occurs 'f' number of trials when a random experiment is repeated for 'n' number of times, then P(A) = $\underset{n \to \propto}{Lt} \frac{f}{n}$

For practical convenience, the above equation may be written as P(A) = $\frac{f}{n}$

Here, probability has between 0 and 1,

i.e. $0 \leq P(A) \leq 1$

Question

The compensation received by 1000 workers in a factory are given in the following table :-

Wages:	80-100	100-120	120-140	140-160	160-180	180-200
No. of Workers:	10	100	400	250	200	40

Find the probability that a worker selected has

(1) Wages under Rs.100/-

(2) Wages above Rs.140/-

(3) Wages between Rs. 120/- and Rs.180/-

Solution

P(that a worker selected has wages under Rs.140/-) = $\frac{10+100+400}{1000} = \frac{510}{1000}$

P(that a worker selected has wages above Rs.140/-=$\frac{250+200+40}{1000} = \frac{490}{1000}$

P(that a worker selected has wages between 120 and 180) = $\frac{400+250+200}{1000} = \frac{850}{1000}$

Subjective (Personalistie) Approach to Probability

The exponents of personalistie approach defines probability as a measure of personal confidence or belief based on whatever evidence is available. For example, if a teacher wants to find out the probability that Mr. X topping in M.Com examination, he may assign a value between zero and one according to his degree of belief for possible occurrence. He may take into account such factors as the past academic performance in terminal examinations etc. and arrive at a probability figure. The probability figure arrived under this method may vary from person to person. Hence it is called subjective method of probability.

Axiomatic Approach (Modern Approach) to Probability

Let 'S' be the sample space of a random experiment, and 'A' be an event of the random experiment, so that 'A' is the subset of 'S'. Then we can associate a real number to the event 'A'. This number will be called probability of 'A' if it satisfies the following three axioms or postulates :-

(1) The probability of an event ranges from 0 and 1.

If the event is certain, its probability shall be 1.

If the event cannot take place, its probability shall be zero.

(2) The sum of probabilities of all sample points of the sample spece is equal to 1. i.e, $P(S) = 1$

(3) If A and B are mutually exclusive (disjoint) events, then the probability of occurrence of either A or B shall be :

$$P(A \cup B) = P(A) + P(B)$$

THEOREMS OF PROBABILITY

There are two important theorems of probability. They are :

1. Addition Theorem
2. Multiplication Theorem

Addition Theorem

Here, there are 2 situations.

(a) Events are mutually exclusive

(b) Events are not mutually exclusive

(a) Addition theorem (Mutually Exclusive Events)

If two events, 'A' and 'B', are mutually exclusive the probability of the occurrence of either 'A' or 'B' is the sum of the individual probability of A and B.

$$P(A \text{ or } B) = P(A) + P(B)$$

i.e., $P(A \cup B) = P(A) + P(B)$

(b)Addition theorem (Not mutually exclusive events)

If two events, A and B are not mutually exclusive the probability of the occurrence of either A or B is the sum of their individual probability minus probability for both to happen.

P(A or B) = P(A) + P(B) – P(A and B)

i.e., $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Question

What is the probability of picking a card that was red or black?

Solution

Here the events are mutually exclusive

P(picking a red card) = $\frac{26}{52}$

P(picking a black card) = $\frac{26}{52}$

∴P (picking a red or black card) = $\frac{26}{52} + \frac{26}{52} = 1$

Question

The probability that a contractor will get a plumping contract is $\frac{2}{3}$ and the probability that he will not get an electric contract is $\frac{5}{9}$. If the probability of getting at least one contract is $\frac{4}{5}$, what is the probability that he will get both the contracts?

Solution

P(getting plumbing contract) = $\frac{2}{3}$

P(not getting electric contract) = $\frac{5}{9}$

P(getting electric contract) = 1 - $\frac{5}{9} = \frac{4}{9}$

P(getting at least one contract) = P (getting electric contract) +

P(getting plumbing contract) - P(getting both)

i.e., $\frac{4}{5} = \frac{4}{9} + \frac{2}{3}$ - P(getting both)

∴P(getting both contracts) = $\frac{4}{9} + \frac{2}{3} - \frac{4}{5}$

$= \frac{20 + 30 - 36}{45} = \frac{14}{45}$

Question

If P(A) = 0.5, P(B) = 0.6, P(A∩B) = 0.2, find:-

(a) P(A∪B)

(b) P(A')

(c) P(A∩B')

(d) P(A'∩B')

Solution

Here the events are not mutually exclusive:-

(a) P(A∪B) = P(A) + P(B) – P(A∩B)

= 0.5 + 0.6 – 0.2

= 0.9

(b) P(A') = 1- P(A)
= 1 – 0.5

=0.5

(c) P(A∩B')
= P(A) – P(A∩B)
= 0.5 – 0.2
= 0.3

(d) P(A'∩B') = 1 – P(A∪B)

= 1- [P(A) + P(B) – P(A∩B)

= 1 – (0.5 + 0.6 – 0.2)

= 1- 0.9

= 0.1

MULTIPLICATION THEOREM

Here there are two situations:

(a) Events are independent

(b) Events are dependent

(a)Multiplication theorem (independent events)

If two events are independent, then the probability of occurring both will be the product of the individual probability

P(A and B) = P(A).P(B)

i.e., P(A∩B) = P(A).P(B)

Question

A bag contains 5 white balls and 8 black balls. One ball is drawn at random from the bag and is then replaced. Again another one is drawn. Find the probability that both the balls are white.

Solution

Here the events are independent

P(drawing white ball in I draw) = $\frac{5}{13}$

P(drawing white ball in II draw) = $\frac{5}{13}$

∴ P(drawing white ball in both draw) = $\frac{5}{13} \times \frac{5}{13}$

$= \frac{25}{169}$

Question

Single coin is tossed for three tones. What is the probability of getting head in all the 3 times?

Solution

P(getting head in all the 3 times) = P(getting H in 1st toss) × P(getting Head in 2nd toss) × P(getting H in 3rd toss)

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

$= \frac{1}{8}$

(b)Multiplication theorem (dependent Events):-

If two events, A and B are dependent, the probability of occurring 2nd event will be affected by the outcome of the first.

P(A∩B) = P(A).P(B/A)

Question

A bag contains 5 white balls and 8 black balls. One ball is drawn at random from the bag. Again, another one is drawn without replacing the first ball. Find the probability that both the balls drawn are white.

Solution

P(drawing a white ball in Ist draw) =$\frac{5}{13}$

P (drawing a white ball in IInd draw) = $\frac{4}{12}$

$= \frac{20}{156}$

Question

The probability that 'A' solves a problem in Maths is $\frac{2}{5}$ and the probability that 'B' solves it is $\frac{3}{8}$. If they try independently find the probability that :-

(i) Both solve the problem.

(ii) at least one solve the problem.

(iii) none solve the problem.

Solution

(i) P(that both solve the problem) = P(that A solves the problem)× P(that B solves the problem)

$= \frac{2}{5} \times \frac{3}{8} = \frac{6}{40} = \frac{3}{20}$

(ii) P(that at least one solve the problem) } = P(that A or B solves the problem)

= P(A solve the problem + P(B solve the problem – P(A and B solve the problem)

$= \frac{2}{5} + \frac{3}{8} - (\frac{2}{5} \times \frac{3}{8})$

$= \frac{2}{5} + \frac{3}{8} - \frac{6}{40}$

$= \frac{16+15-6}{40}$

$= \frac{25}{40} = \frac{5}{8}$

(iii) P(that none solve the problem) =1-P(at least one solve the problem)

= 1 – P(A or B solve the problem)

=1- [P(A solve the problem + P(B solve the problem) - P(A & B solve the problem)]

$= 1- \left[\frac{2}{5} + \frac{3}{8} - (\frac{2}{5} \times \frac{3}{8})\right]$

$= 1-(\frac{2}{5} + \frac{3}{8} - \frac{6}{40})$

$$= 1 - \left(\frac{16+15-6}{40}\right)$$

$$= 1 - \frac{25}{40} = \frac{15}{40} = \frac{3}{8}$$

Question

A university has to select an examiner from a list of 50 persons. 20 of them are women and 30 men. 10 of them know Hindi and 40 do not. 15 of them are teachers and remaining are not. What is the probability that the university selecting a Hindi knowing woman teacher?

Solution

Here the events are independent.

P(selecting Hindi knowing Woman teacher) = P(selecting Hindi knowing person, woman and teacher

P(selecting Hindi Knowing persons) $= \frac{10}{50}$

P(selecting woman) $= \frac{20}{50}$

P (selecting teacher) $= \frac{15}{50}$

P(selection Hindi knowing Woman teacher $= \frac{10}{50} \times \frac{20}{50} \times \frac{15}{50}$

$$= \frac{2}{10} \times \frac{4}{10} \times \frac{3}{10} = \frac{24}{1000}$$

$$= \frac{3}{125}$$

Question

'A' speaks truth in 70%cases and 'B' in 85% cases. In what percentage of cases they likely to contradict each other in stating the same fact?

Let P(A) = Probability that 'A' speaks truth.

P(A') = Probability that 'A' does not speak truth

P(B) = Probability that 'B' speaks truth

P(B') = Probability that 'B' does not speak truth

P(A) = 70% = 0.7

P(A') =30% =0.3

P(B) = 85% = 0.85

P(B') = 15% = 0.15

∴ P (A and B contradict each other) = P('A' speaks truth and 'B' does not OR, A does not speak truth & B speaks

=P(A & B') ∪(A' &B)

=(0.7 ×0.15) + (0.3 ×0.85)

=0.105 + 0.255

= 0.360

∴ Percentage of cases in in which A and B contradict each other = 0.360 ×100

= 36 %

Question

20% of students in a university are graduates and 80% are undergraduates. The probability that graduate student is married is 0.50 and the probability that an undergraduate student is married is 0.10. If one student is selected at random, what is the probability that the student selected is married?

P(selecting a married student) = P(selecting a graduate married student or selecting an undergraduate married student)

= P(Selecting a graduate & married OR selecting un undergraduate & married)

= (20% ×0.50) + (80% ×0.10)

= $(\frac{20}{100} \times 0.50) + (\frac{80}{100} \times 0.10)$

= (0.2 × 0.50) + (0.8 × 0.1)

= 0.1 +0.08

= 0.18

Question

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

Solution

P(that new product will be introduced) = P(that new product is introduced by first set OR the new product is introduced by second set)

= P (I^{st} set wins & I^{st} introduced the new produced OR II^{nd} set wins the new product)

= $(0.6 \times 0.8) + (0.4 \times 0.3)$

= $0.48 + 0.12$

= $\underline{0.60}$

Question

A certain player say Mr. X is known to win with possibility 0.3 if the truck is fast and 0.4 if the track is slow. For Monday there is a 0.7 probability of a fast track and 0.3 probability of a slow track. What is the probability that Mr. X will win on Monday?

Solution

P(X will won on Monday) = P(to win in fast track OR to win in slow track)

= P (to get fast track & to win OR to get slow track & to win)

= $(0.7 \times 0.3) + (0.3 \times 0.4)$

= $\underline{0.21} + 0.12$

= $\underline{0.33}$

CONDITIONAL PROBABILITY

Multiplication theorem states that if two events, A and B, are dependent events then, the probability of happening both will be the product of P(A) and P(B/A).

i.e., P(A and B) or $P(A \cap B)$ = P(A).P(B/A)

Here, P (B/A) is called Conditional probability

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

$\text{i.e., } P(A).P(B/A) = P(A \cap B)$

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

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

If 3 event, A, B and C and dependent events, then the probability of happening A, B and C is :-

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

i.e., $P(A).P(B/A).P(C/AB) = P(A \cap B \cap C)$

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

Question

If $P(A) = \frac{1}{13}$, $P(B) = \frac{1}{4}$ and $P(A \cap B) = 1/52$, find:-

(a) P(A/B)

(b) P(B/A)

Solution

Here we know the events are dependent

(a) $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{1}{4}} = \frac{1}{52} \times \frac{4}{1} = \frac{4}{52} = \frac{1}{13}$

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

$= \frac{\frac{1}{52}}{\frac{1}{13}} = \frac{1}{52} \times \frac{13}{1} = \frac{13}{52} = \frac{1}{4}$

Inverse Probability

If an event has happened as a result of several causes, then we may be interested to find out the probability of a particular cause of happening that events. This type of problem is called inverse probability.

Baye's theorem is based upon inverse probability.

BAYE'S THEOREM:

Baye's theorem is based on the proposition that probabilities should revised on the basis of all the available information. The revision of probabilities based on available information will help to reduce the risk involved in decision-making. The probabilities before revision is called priori probabilities and the probabilities after revision is called posterior probabilities.

According to Baye's theorem, the posterior probability of event (A) for a particular result of an investigation (B) may be found from the following formula:-

$$P(A/B) = \frac{P(A).P(B)}{P(A).P(B) + P(Not\ A).P(\frac{B}{Not}A)}$$

Steps in computation

1. Find the prior probability
2. Find the conditional probability.
3. Find the joint probability by multiplying step 1 and step 2.
4. Find posterior probability as percentage of total joint probability.

Question

A manufacturing firm produces units of products in 4 plants, A, B, C and D. From the past records of the proportions of defectives produced at each plant, the following conditional probabilities are set:-

A: 0.5; B: 0.10; C:0.15 and D:0.02

The first plant produces 30% of the units of the output, the second plant produces 25%, third 40% and the forth 5%

A unit of the products made at one of these plants is tested and is found to be defective. What is the probability that the unit was produced in Plant C.

Solution

Computation of Posterior probabilities				
Machine	Priori Probability	Conditional Probabilities	Joint Probability	Posterior Probability
A	0.30	0.05	0.015	$\frac{0.015}{0.101}$ = 0.1485
B	0.25	0.10	0.025	$\frac{0.025}{0.101}$= 0.2475
C	0.40	0.15	0.060	$\frac{0.060}{0.101}$=0.5941
D	0.05	0.02	0.001	$\frac{0.001}{0.101}$=0.0099
			0.101 ====	1.0000 =====

Probability that defective unit was produced in Machine C = 0.5941

Question

In a bolt manufacturing company machine I, II and III manufacture respectively 25%, 35% and 40%. Of the total of their output, 5%, 4% and 2% are defective bolts. A bolt is drawn at random from the products and is found to be defective. What are the probability that it was manufactured by :-

(a)Machine I

(b) Machine II

(c)Machine III

Solution

Computation of Posterior probabilities				
Machine	Priori Probability	Conditional Probabilities	Joint Probability	Posterior Probability
I	0.25	0.05	0.0125	0.362
II	0.35	0.04	0.0140	0.406
III	0.40	0.02	0.0080	0.232
			0.0345	1.000

P(that the bolt was manufactured by Machine I) = 0.362

P(that the bolt was manufactured by Machine II) = 0.406

P(that the bolt was manufactured by Machine III) = 0.232

Question

The probability that a doctor will diagnose a particular disease correctly is 0.6. The probability that a patient will die by his treatment after correct diagnosis is 0.4 and the probability of death by wrong diagnosis is 0.7. A patient of the doctor who had the disease died. What is the probability that his disease was not correctly diagnosed?

Solution

Computation of Posterior probabilities				
Nature of Diagnosis	Priori Probability	Conditional Probabilities	Joint Probability	Posterior Probability
Correct	0.6	0.4	0.24	$\frac{0.24}{0.52} = 0.462$
Not correct	0.4	0.7	0.28	$\frac{0.28}{0.52} = 0.538$
			0.52	1.000

Probability that the disease was not correctly diagnosed = 0.538

Question

There are two Urns, one containing 5 white balls and 4 black balls; and the other containing 6 white balls and 5 black balls. One Urn is chosen and one ball is drawn. If it is white, what is the probability that the Urn selected is the first?

Solution

Computation of Posterior probabilities				
No. of Urn	Probability of drawing white ball (Priori Probability)	Conditional Probabilities	Joint Probability	Posterior Probability
Ind	$\frac{5}{9}$	$\frac{1}{2}$	$\frac{5}{18}$ = 0.2778	$\frac{0.2778}{0.5505}$= 0.5046
IInd	$\frac{6}{11}$	$\frac{1}{2}$	$\frac{6}{22}$= 0.2727	$\frac{0.2727}{0.5505}$= 0.4954
			0.5505	1.00

P(that the white balls drawn is from Urnn I = 0.5046

=====

CHAPTER - 5

PROBABILITY DISTRIBUTION (THEORETICAL DISTRIBUTION)

DEFINITION

Probability distribution (Theoretical Distribution) can be defined as a distribution obtained for a random variable on the basis of a mathematical model. It is obtained not on the basis of actual observation or experiments, but on the basis of probability law.

Random variable

Random variable is a variable who value is determined by the outcome of a random experiment. Random variable is also called chance variable or stochastic variable.

For example, suppose we toss a coin. Obtaining of head in this random experiment is a random variable. Here the random variable of "obtaining heads" can take the numerical values.

Now, we can prepare a table showing the values of the random variable and corresponding probabilities. This is called probability distributions or theoretical distribution.

In the above, example probability distribution is :-

Obtaining of heads X	Probability of obtaining heads P(X)
0	$1/2$
1	$1/2$
	$\sum P(X) = 1$

Properties of Probability Distributions:

1. Every value of probability of random variable will be greater than or equal to zero.
   i.e., $P(X) \geq 0$
   i.e., $P(X) \neq$ Negative value

2. Sum of all the probability values will be 1
   $\sum P(X) = 1$

Question

A distribution is given below. State whether this distribution is a probability distribution.

X:	0	1	2	3	4
P(X):	0.01	0.10	0.50	0.30	0.90

Solution

Here all values of P(X) are more that zero; and sum of all P(X) value is equal to 1

Since two conditions, namely P(X) ≥0 and ∑P(X) = 1, are satisfied, the given distribution is a probability distribution.

MATHEMATICAL EXPECTATION

(EXPECTED VALUE)

If X is a random variable assuming values x_1, x_2, x_3,……,$x_n$with corresponding probabilities P_1, P_2, P_3,……,P_n, then the operation of X is defined as X_1P_1+ X_2P_2+ X_3P_3+……+ X_nP_n.

$$E(X) = \sum[X.P(X)]$$

Question

A petrol pump proprietor sells on an average Rs. 80,000/- worth of petrol on rainy days and an average of Rs. 95.000 on clear days. Statistics from the meteorological department show that the probability is 0.76 for clear weather and 0.24 for rainy weather on coming Wednesday. Find the expected value of petrol sale on coming Wednesday.

Expected Value E(X) $\} = \sum[X.P(X)]$

$= (80.000 \times 0.24)+(95000 \times 0.76)$

$= 19.200 + 72.200$

$=$ Rs. 91,400

Question

There are three alternative proposals before a business man to start a new project:-

Proposal I: Profit of Rs. 5 lakhs with a probability of 0.6 or a loss of Rs. 80,000 with a probability of 0.4.

Proposal II: Profit of Rs. 10 laksh with a probability of 0.4 or a loss of Rs. 2 lakhs with a probability of 0.6

Proposal III: Profit of Rs. 4.5 lakhs with a probability of 0.8 or a loss of Rs. 50,000 with a probability of 0.2

If he wants to maximize profit and minimize the loss, which proposal he should prefer?

Solution

Here, we should calculate the mathematical expectation of each proposal.

Expected Value E(X) = $\sum[X.P(X)]$

Expected value of Proposal I } = (5,00.000 × 0.6) +(80.000 × 0.4)

= 3,00.000 – 32,000

= Rs. 2,68,000

======

Expected value of Proposal II } = (10,00.000 × 0.4) +(-2,00.000 × 0.6)

= 4,00.000 – 1,20.000

= 2,80,000.

======

Expected value of Proposal III } = (4,50,000 × 0.8) +(-50,000 × 0.2)

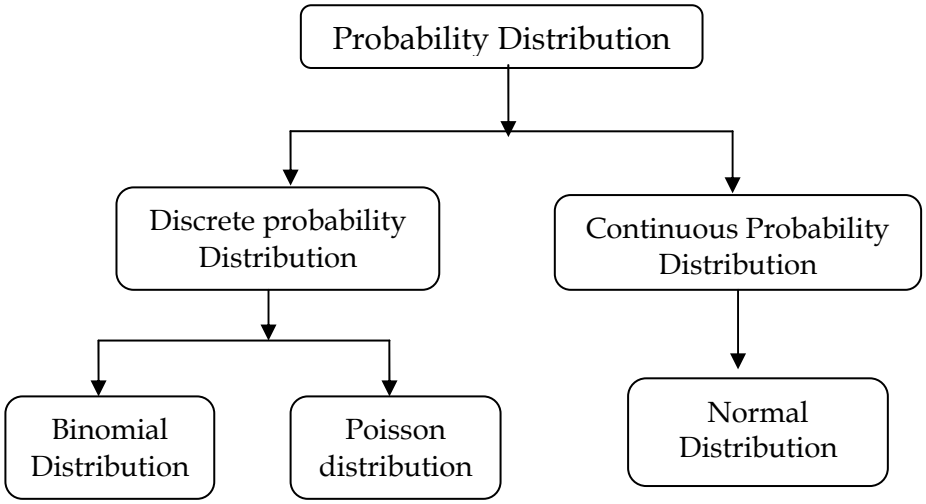
= 3,60,000 – 10,000

= 3,50,000

======

Since expected value is highest in case of proposal III, the businessman should prefer the proposal III.

Classification of Probability Distribution



Discrete Probability Distribution

If the random variable of a probability distribution assumes specific values only, it is called discrete probability distributions. Binomial distribution and poisson distribution are discrete probability distributions.

Continuous Probability Distributions:-

If the random variable of a probability distribution assumes any value in a given interval, then it is called continuous probability distributions. Normal distributions is a continuous probability distribution.

CHAPTER - 6

BIONOMIAL DISTRIBUTION

Meaning & Definition:

Binomial Distribution is associated with James Bernoulli, a Swiss Mathematician. Therefore, it is also called Bernoulli distribution. Binomial distribution is the probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure. In other words, it is used to determine the probability of success in experiments on which there are only two mutually exclusive outcomes. Binomial distribution is discrete probability distribution.

Binomial Distribution can be defined as follows: "A random variable r is said to follow Binomial Distribution with parameters n and p if its probability function is:

$$P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$$

Where, P = probability of success in a single trial

q = 1 – p

n = number of trials

r = number of success in 'n' trials.

Assumption of Binomial Didstribution OR

(Situations where Binomial Distribution can be applied)

Binomial distribution can be applied when:-

1. The random experiment has two outcomes i.e., success and failure.
2. The probability of success in a single trial remains constant from trial to trial of the experiment.
3. The experiment is repeated for finite number of times.
4. The trials are independent.

Properties (features) of Binomial Distribution:

1. It is a discrete probability distribution.
2. The shape and location of Binomial distribution changes as 'p' changes for a given 'n'.
3. The mode of the Binomial distribution is equal to the value of 'r' which has the largest probability.
4. Mean of the Binomial distribution increases as 'n' increases with 'p' remaining constant.
5. The mean of Binomial distribution is np.
6. The Standard deviation of Binomial distribution is $\sqrt{npq}$
7. If 'n' is large and if neither 'p' nor 'q' is too close zero, Binomial distribution may be approximated to Normal Distribution.
8. If two independent random variables follow Binomial distribution, their sum also follows Binomial distribution.

Qn: Six coins are tossed simultaneously. What is the probability of obtaining 4 heads?

Sol: $P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$

$r = 4$

$n = 6$

$p = ½$

$q = 1 - p = 1 - ½ = ½$

$\therefore p(r = 4) = {}^{6}C_{4} (½)^{4} (½)^{6-4}$

$= \frac{6!}{(6-4)!4!} \times (½)^{4+2}$

$= \frac{6!}{2!4!} \times (½)^{6}$

$= \frac{6 X 5}{2 X 1} \times \frac{1}{64}$

$= \frac{30}{128}$

$= \underline{0.234}$

Qn: The probability that Sachin Tendulkar scores a century in a cricket match is $\frac{1}{3}$. What is the probability that out of 5 matches, he may score century in : -

(1) Exactly 2 matches
(2) No match

Sol: Here $p = \frac{1}{3}$

$q = 1 - \frac{1}{3} = \frac{2}{3}$

$P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$

Probability that Sachin scores centuary in exactly 2 matches is:

$p(r = 2) = {}^{5}C_{2} \left(\frac{1}{3}\right)^{2} \left(\frac{2}{3}\right)^{3}$

$= \frac{5!}{(5-2)!2!} \times \frac{1}{9} \times \frac{8}{27}$

$= \frac{5 X 4}{2 X 1} \times \frac{1}{9} \times \frac{8}{27}$

$= \frac{160}{486}$

$= \frac{80}{243} = \underline{0.329}$

(1) Probability that Sachin scores centuary in no matches is:

$p(r = 0) = {}^{5}C_{0} \left(\frac{1}{3}\right)^{0} \left(\frac{2}{3}\right)^{5-0}$

$= \frac{5!}{(5-0)!0!} \text{ x } 1 \qquad \text{x } \left(\frac{2}{3}\right)^5$

$= \frac{5!}{5!X\ 0!} \text{ x } 1 \text{ x } \frac{32}{243}$

$= \frac{32}{243}$

$= \underline{0.132}$

Qn: Consider families with 4 children each. What percentage of families would you expect to have :-

(a) Two boys and two girls
(b) At least one boy
(c) No girls
(d) At the most two girls

Sol: p (having a boy) $= \frac{1}{2}$

p (having a girl) $= \frac{1}{2}$

n = 4

a) P (getting 2 boys & 2 girls) = p (getting 2 boys) = p (r = 2)

$P(r) = {}^{n}C_{r}\, p^r q^{n-r}$

$p(r=2) = {}^{4}C_{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-2}$

$= \frac{4!}{(4-2)!2!} \text{ x } \left(\frac{1}{2}\right)^2 \text{X} \left(\frac{1}{2}\right)^2$

$= \frac{4!}{2!2!} \text{ x } \left(\frac{1}{2}\right)^{2+2}$

$= \frac{4\ X\ 3}{2\ X\ 1} \text{X} \left(\frac{1}{2}\right)^4$

$= \frac{12}{2} \text{ X } \frac{1}{16}$

$= \frac{12}{32} = \frac{3}{8}$

$\therefore$% of families with 2 boys & 2 girls $= \frac{3}{8} \text{ x } 100$

$= 37.5\%$

b) Probability of having at least one boy

= p (having one boy or having 2 boys or having 3 boys or having 4 boys)

= p (having one boy) + p (having 2 boys) + p (having 3 boys)
+ p (having 4 boys)

= p (r=1) + p (r = 2) + p(r = 3) + p (r = 4)

$P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$

$p\,(r = 1) = {}^{4}C_{1} \left(\frac{1}{2}\right)^{1} \left(\frac{1}{2}\right)^{4-1}$

$= 4 \text{ X } \left(\frac{1}{2}\right)^{4} = 4 \text{ X } \frac{1}{16} = \frac{4}{16}$

$p\,(r = 2) = {}^{4}C_{2} \left(\frac{1}{2}\right)^{2} \left(\frac{1}{2}\right)^{4-2}$

$= 6 \text{ X } \frac{1}{16} = \frac{6}{16}$

$p\,(r = 3) = {}^{4}C_{3} \left(\frac{1}{2}\right)^{3} \left(\frac{1}{2}\right)^{4-3}$

$= 4 \text{ X } \left(\frac{1}{2}\right)^{4} = 4 \text{ X } \frac{1}{16} = \frac{4}{16}$

$p\,(r = 4) = {}^{4}C_{4} \left(\frac{1}{2}\right)^{4} \left(\frac{1}{2}\right)^{4-4}$

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

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

$= 1 \text{ X } \frac{1}{16} = \frac{1}{16}$

$\therefore p(r = 1 \text{ or } r=2 \text{ or } r=3 \text{ or } r=4) = \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$

$= \frac{15}{16}$

$\therefore$ % of families with at least one boy $= \frac{15}{16} \text{ x } 100 = 93.75\%$

c) Probability of having no girls = p (having 4 boys)

$p\,(r = 4) = {}^{4}C_{4} \left(\frac{1}{2}\right)^{4} \left(\frac{1}{2}\right)^{4-4}$

$= \frac{4!}{(4-4)!4!} \text{ x } \left(\frac{1}{2}\right)^{4+0}$

$= \frac{4!}{0!4!} \text{ x } \frac{1}{16}$

$= \frac{1}{16}$

$\therefore$ % of families with at least no girls $= \frac{1}{16} \text{ x } 100 = 6.25\%$

d) Probability of having at the most 2 girls

= p (having 2 girls or having 1 girl or having no girl)
= p (having 2 boys or having 3 boys or having 4 boys)
= p (r =2) + p (r =3) + p (r = 4)

$p\,(r = 2) = {}^{4}C_{2} \left(\frac{1}{2}\right)^{2} \left(\frac{1}{2}\right)^{4-2}$

$= \frac{6}{16}$

$p(r = 3) = {}^4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{4-3}$

$= \frac{4}{16}$

$p(r = 4) = {}^4C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{4-4}$

$= {}^4C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0$

$= 1 \text{ X } \left(\frac{1}{2}\right)^4$

$= \frac{1}{16}$

$\therefore$ p(having at the most 2 girls) = $\frac{6}{16} + \frac{4}{16} + \frac{1}{16}$

$= \frac{11}{16}$

$\therefore$ % of families with at the most 2 girls = $\frac{11}{16}$ x 100 = 68.75%

Mean and Standard Deviation of Binomial Distribution

Mean of Binomial Distribution = np

Standard Deviation of Binomial Distribution = $\sqrt{npq}$

Qn: For a Binomial Distribution, mean = 4 and variance = $\frac{12}{9}$. Find n.

Sol: Mean = np = 4

Standard Deviation = $\sqrt{npq}$

$\therefore$ Variance = Standard Deviation ^{2}

$= \left(\sqrt{npq}\right)^2$

= npq

npq = $\frac{12}{9}$

Divide npq by np to get the value of q

i.e. $\frac{npq}{np}$ = q

q = $\frac{npq}{np} = \frac{\frac{12}{9}}{\frac{4}{1}}$

$= \frac{12}{9} \text{ X } \frac{1}{4} = \frac{3}{9} = \frac{1}{3}$

$q = \frac{1}{3}$

$p = 1 - q$

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

$np = 4$

$n \times \frac{2}{3} = 4$

$n = 4 \div \frac{2}{3}$

$n = 4 \times \frac{3}{2} = 6$

Qn: For a Binomial Distribution, mean is 6 and Standard Deviation is $\sqrt{2}$. Find the parameters.

Sol: Mean (np) = 6

Standard Deviation ($\sqrt{npq}$) = $\sqrt{2}$

$\therefore$ npq = 2

$\frac{npq}{np} = \frac{2}{6}$

$q = \frac{1}{3}$

$\therefore p = 1 - q$

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

$np = 6$

$n \times \frac{2}{3} = 6$

$\therefore n = \frac{6}{\frac{2}{3}} = 6 \times \frac{3}{2} = 9$

Value of parameters:

$p = \frac{2}{3}$ $q = \frac{1}{3}$ $n = 9$

Qn: In a Binomial Distribution consisting 5 independent trials, probability for 1 and 2 successes are 0.4096 and 0.2048 respectively. Find the parameter p.

Sol: As there are 5 trials, the terms of the Binomial Distribution are:-

p (r = 0) ; p (r = 1); p (r = 2); p (r = 3); p (r = 4) and p (r = 5)

p (r = 1) = 0.4096

p (r = 2) = 0.2048

$p(r = 1) = {}^{n}C_{r}\, p^{r}q^{n-r} = {}^{5}C_{1}\, p^{1}q^{5-1} = 5pq^{4} = 0.4096$

$p(r = 2) = {}^{n}C_{r}\, p^{r}q^{n-r} = {}^{5}C_{2}\, p^{2}q^{5-2} = 10p^{2}q^{3} = 0.2048$

Divide the first term by the second term

$$\frac{5pq^4}{10\,p^2\,q^3} = \frac{0.4096}{0.2048}$$

i.e., $\frac{q}{2p} = \frac{2}{1}$

$4p = q$

$4p = 1 - p$

$4p + p = 1$

$5p = 1$

$p = \frac{1}{5}$

Fitting a Binomial Distribution

Steps:

1. Find the value of n, p and q
2. Substitute the values of n, p and q in the Binomial Distribution function of ${}^{n}C_{r}\, p^{r}q^{n-r}$
3. Put r = 0, 1, 2, …….. in the function ${}^{n}C_{r}\, p^{r}q^{n-r}$
4. Multiply each such terms by total frequency (N) to obtain the expected frequency.

Qn: Eight coins were tossed together for 256 times. Fit a Binomial Distribution of getting heads. Also find mean and standard deviation.

Sol: p (getting head) = $p = \frac{1}{2}$

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

$n = 8$

Binomial Distribution function is $p(r) = {}^{n}C_{r}\, p^{r}q^{n-r}$

Put r= 0, 1, 2, 3 …….. 8, then are get the terms of the Binomial Distribution.

No. of heads i.e. r	P (r)	Expected Frequency P (r) x N N = 256
0	$^8C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^8 = 1 \times 1 \times \frac{1}{256} = \frac{1}{256}$	1
1	$^8C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^7 = 8 \times \frac{1}{256} = \frac{8}{256}$	8
2	$^8C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^6 = 28 \times \frac{1}{256} = \frac{28}{256}$	28
3	$^8C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^5 = 56 \times \frac{1}{256} = \frac{56}{256}$	56
4	$^8C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^4 = 70 \times \frac{1}{256} = \frac{70}{256}$	70
5	$^8C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^3 = 56 \times \frac{1}{256} = \frac{56}{256}$	56
6	$^8C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^2 = 28 \times \frac{1}{256} = \frac{28}{256}$	28
7	$^8C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^1 = 8 \times \frac{1}{256} = \frac{8}{256}$	8
8	$^8C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^0 = 1 \times \frac{1}{256} = \frac{1}{256}$	1

Mean = np

$= 8 \times \frac{1}{2} = 4$

Standard Deviation = $\sqrt{npq}$

$= \sqrt{8 \times \frac{1}{2} \times \frac{1}{2}} \quad = \sqrt{2} \quad = \underline{1.4142}$

CHAPTER - 7

POISSON DISTRIBUTION

Meaning and Definition:

Poisson Distribution is a limiting form of Binomial Distribution. In Binomial Distribution, the total number of trials are known previously. But in certain real life situations, it may be impossible to count the total number of times a particular event occurs or does not occur. In such cases Poisson Distribution is more suitable.

Poison Distribution is a discrete probability distribution. It was originated by Simeon Denis Poisson.

The Poisson Distribution is defined as:-

$$p(r) = \frac{e^{-m}.m^r}{r!}$$

Where r = random variable (i.e., number of success in '**n**' trials.

e = 2.7183

m = mean of poisson distribution

Properties of Poisson Distribution

1. Poisson Distribution is a discrete probability distribution.
2. Poisson Distribution has a single parameter 'm'. When 'm' is known all the terms can be found out.
3. It is a positively skewed distribution.
4. Mean and Varriance of Poisson Distribution are equal to 'm'.
5. In Poisson Distribution, the number of success is relatively small.
6. The standard deviation of Poisson Distribution is $\sqrt{m}$.

Practical situations where Poisson Distribution can be used

1. To count the number of telephone calls arising at a telephone switch board in a unit of time.
2. To count the number of customers arising at the super market in a unit of time.
3. To count the number of defects in Statistical Quality Control.
4. To count the number of bacterias per unit.
5. To count the number of defectives in a park of manufactured goods.
6. To count the number of persons dying due to heart attack in a year.
7. To count the number of accidents taking place in a day on a busy road.

Qn: A fruit seller, from his past experience, knows that 3% of apples in each basket will be defectives. What is the probability that exactly 4 apples will be defective in a given basket?

Sol: $p(r) = \frac{e^{-m} . m^r}{r!}$

$m = 3$

$\therefore p(r = 4) = \frac{e^{-3} . 3^4}{4!} = \frac{0.04979 \; x \; 81}{4 \; x \; 3 \; x \; 2 \; x \; 1}$

$= \frac{0.04979 \; x \; 81}{24}$

$= 0.16804$

Qn: It is known from the past experience that in a certain plant, there are on an average four industrial accidents per year. Find the probability that in a given year there will be less than four accidents. Assume poisson distribution.

Sol: $p(r<4) = p(r = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3)$

$= p(r = 0) + p(r = 1) + p(r = 2) + p(r = 3)$

$P(r) = \frac{e^{-m} . m^r}{r!}$

$m = 4$

$\therefore p(r = 0) = \frac{e^{-4} . 4^0}{0!} = \frac{0.01832 \; x \; 1}{1} = 0.01832$

$p(r = 1) = \frac{e^{-4} . 4^1}{1!} = \frac{0.01832 \; x \; 4}{1} = 0.07328$

$p(r = 2) = \frac{e^{-4} . 4^2}{2!} = \frac{0.01832 \; x \; 16}{2 \; x \; 1} = 0.14656$

$p(r = 3) = \frac{e^{-4} . 4^3}{3!} = \frac{0.01832 \; x \; 64}{3 \; x \; 2 \; x \; 1} = 0.19541$

$\therefore p(r < 4) = 0.01832 + 0.07328 + 0.14656 + 0.19541$

$= 0.43357$

Qn: Out of 500 items selected for inspection, 0.2% are found to be defective. Find how many lots will contain exactly no defective if there are 1000 lots.

Sol: $p = 0.2\% = 0.002$

$n = 500$

$m = np = 500 \text{ x } 0.002 = 1$

$p(r) = \frac{e^{-m} . m^r}{r!}$

$p(r = 0) = \frac{e^{-1} . 1^0}{0!} = \frac{0.36788 \times 1}{1} = 0.36788$

∴ Number of lots containing no defectives if there are 1000 lots = 0.36788 x 1000

= 367.88

= 368

Qn: In a factory manufacturing optical lenses, there is a small chance of $\frac{1}{1500}$ for any one lense to be defective. The lenses are supplied in packets of 10. Use Poisson Distribution to calculate the approximate number of packets containing (1) one defective (2) no defective in a consignment of 20,000 packets. You are given that $e^{-0.02} = 0.9802$.

Sol: n = 10

p = probability of manufacturing defective lense = $\frac{1}{500} = 0.002$

m = np = 10 x 0.002 = 0.02

$p(r) = \frac{e^{-m} . m^r}{r!}$

$p(r = 1) = \frac{e^{-0.02} \times 0.02^1}{1!} = \frac{0.9802 \times 0.02}{1} = 0.019604$

∴ No. of packets containing one defective lense = 0.019604 x 20,000

= 392

$p(r = 0) = \frac{e^{-0.02} \times 0.02^0}{0!} = \frac{0.9802 \times 1}{0} = 0.9802$

∴ No. of packets containing no defective lense = 0.9892 x 20,000

= 19604

Qn: A Systematic sample of 100 pages was taken from a dictionary and the observed frequency distribution of foreign words per page was found to be as follows:

No. of foreign words per page (x) :	0	1	2	3	4	5	6
Frequency (f) :	48	27	12	7	4	1	1

Calculate the expected frequencies using Poisson Distribution.

Sol: $p(r) = \frac{e^{-m} . m^r}{r!}$

Here first we have to find out 'm'.

Computation of mean (m)		
x	f	fx
0	48	0
1	27	27
2	12	24
3	7	21
4	4	16
5	1	5
6	1	6
	N = 100	Σfx = 99

$\bar{X} = \frac{\Sigma fx}{N} = \frac{99}{100} = 0.99$

$\therefore$ m = 0.99

$\therefore$ Poisson Distribution = $\frac{e^{-0.99}\, x\, (0.99)^r}{r!}$

Computation of expected frequencies		
x	**p (x)**	**Expected frequency Nx p(x)**
0	$\frac{e^{-0.99}\, x\, (0.99)^0}{0!} = 0.3716$	100 x 0.3716 = 37.2
1	$\frac{e^{-0.99}\, x\, (0.99)^1}{1!} = 0.3679$	100 x 0.3679 = 36.8
2	$\frac{e^{-0.99}\, x\, (0.99)^2}{2!} = 0.1821$	100 x 0.1821 = 18.21
3	$\frac{e^{-0.99}\, x\, (0.99)^3}{3!} = 0.0601$	100 x 0.0601 = 6
4	$\frac{e^{-0.99}\, x\, (0.99)^4}{4!} = 0.0149$	100 x 0.0149 = 1.5
5	$\frac{e^{-0.99}\, x\, (0.99)^5}{5!} = 0.0029$	100 x 0.0029 = 0.3
6	$\frac{e^{-0.99}\, x\, (0.99)^6}{6!} = 0.0005$	100 x 0.0005 = 0.1

Hence, the expected frequencies of this Poisson Distribution are:-

No. of foreign words per page :	0	1	2	3	4	5	6
Expected frequencies (Rounded):	37	37	18	6	2	0	0

CHAPTER - 8

NORMAL DISTRIBUTION

The normal distribution is a continuous probability distribution. It was first developed by De-Moivre in 1733 as limiting form of binomial distribution. Fundamental importance of normal distribution is that many populations seem to follow approximately a pattern of distribution as described by normal distribution. Numerous phenomena such as the age distribution of any species, height of adult persons, intelligent test scores of students, etc. are considered to be normally distributed.

Definition of Normal Distribution

A continuous random variable 'X' is said to follow Normal Distribution if its probability function is:

$$P(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \times e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

π = 3.146

e = 2.71828

μ = mean of the distribution

σ = standard deviation of the distribution

Properties of Normal Distribution (Normal Curve)

1. Normal distribution is a continuous distribution.
2. Normal curve is symmetrical about the mean.
3. Both sides of normal curve coincide exactly.
4. Normal curve is a bell shaped curve.
5. Mean, Median and Mode coincide at the centre of the curve.
6. Quantities are equi-distant from median.
   $Q_3 - Q_2 = Q_2 - Q_1$
7. Normal curve is asymptotic to the base line.
8. Total area under a normal curve is 100%.
9. The ordinate at the mean divide the whole area under a normal curve into two equal parts. (50% on either side).
10. The height of normal curve is at its maximum at the mean.
11. The normal curve is unimodel, i.e., it has only one mode.
12. Normal curve is mesokurtic.
13. No portion of normal curve lies below the x-axis.
14. Theoretically, the range of normal curve is $-\alpha$ to $+\alpha$. But practically the range is $\mu - 3\sigma$ to $\mu + 3\sigma$.

$\mu \pm 1\sigma$ covers 68.27% area

$\mu \pm 2\sigma$ covers 95.45% area

$\mu \pm 3\sigma$ covers 98.73% area.

Importance (or uses) of Normal Distribution

The normal distribution is of central importance in statistical analysis because of the following reasons:-

1. The discrete probability distributions such as Binomial Distribution and Poisson Distribution tend to normal distribution as 'n' becomes large.
2. Almost all sampling distributions conform to the normal distribution for large values of 'n'.
3. Many tests of significance are based on the assumption that the parent population from which samples are drawn follows normal distribution.
4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. Normal distribution finds applications in Statistical Quality Control.
6. Many distributions in social and economic data are approximately normal. For example, birth, death, etc. are normally distributed.

Area under Standard Normal Curve

In case of normal distribution, probability is determined on the basis of area. But to find out the area we have to calculate the ordinate of 2 – scale.

The scale to which the standard deviation is attached is called 2-scale.

$$Z = \frac{\quad}{\quad}$$

Qn: Find $p(z > 1.8)$

Sol: $z > 1.8$ means the area above 1.8; i.e., the area to the right of 1.8

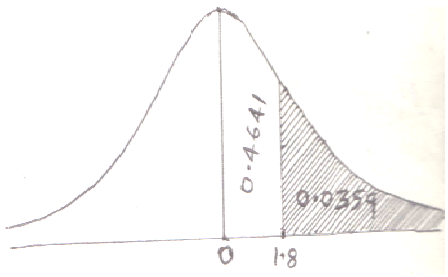
Area upto 1.8 (Table value of 1.8) = 0.4641

Total area on the right side = 0.5

$\therefore$ Area to the right of 1.8 = 0.5 – 0.4641

= 0.0359

$\therefore$ $p(z > 1.8) = 0.0359$



Qn: Find p(z < -1.5)

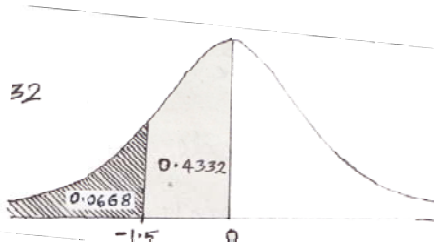
Sol: z < -1.5 means the area to the left of -1.5

Area between o and -1.5 (Table value of 1.5) = 0.4332

Total area on the left side = 0.5

∴ Area to the left -1.5 = 0.5 – 0.4332

= 0.0668



Qn: Find p(z < 1.96)

Sol: z < 1.96 means the entire area to the left of +1.96

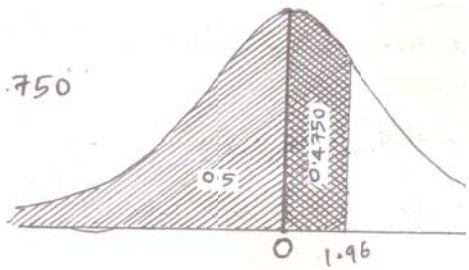
Table value of 1.96 (Area between 0 and 1.96) = 0.4750

Total area on the left side of normal curve = 0.5

∴ Area to the left of 1.96 = 0.4750 + 0.5

= 0.9750

∴ p(z < 1.96) = 0.9750



Qn: Find p(-1.78 < z < 1.78)

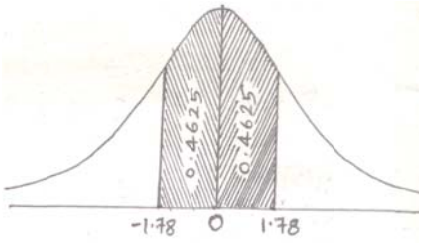
Sol: -1.78 < z < 1.78 means the entire area between -1.78 and +1.78

Table value of 1.78 (Area between 0 and 1.78) = 0.4625

∴ Area between -1.78 and +1.78 = 0.4625 + 0.4625

= 0.9250

∴ p(-1.78 < z < 1.78) = 0.9250



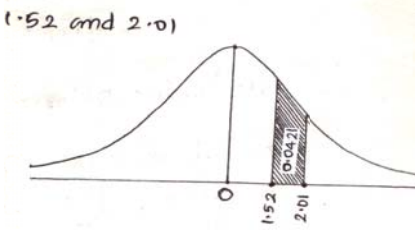
Qn: Find p (1.52 < z < 2.01)

Sol: 1.52 < z < 2.01 means area between 1.52 and 2.01

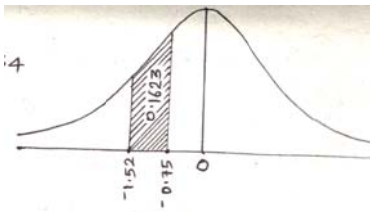
Table value of 1.52 (Area between 0 and 1.52) = 0.4357

Table value of 2.01 (Area between 0 and 2.01) = 0.4778

∴ Area between 2.01 and 1.52 = 0.4778 – 0.4357 = 0.0421



Qn: Find p (-1.52 < z < -0.75)



Sol: -1.52 < z < -0.75 means area between -1.52 and -0.75

Table value of 0.75 (Area between 0 and -0.75) = 0.2734

Table value of 1.52 (Area between 0 and -1.52) = 0.4357

∴ Area between -0.75 and -1.52 = 0.4357 – 0.2734

= 0.1623

p (-1.52 < z < -0.75) = 0.1623

Qn: Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches. How many soldiers of a regiment of 1000 would you expect to be over six feet tall?

Sol: 6 feet = 12 x 6 = 72 inches

Given μ = 68.22 inches

Varriance = 10.8 inches

$\therefore \sigma = \sqrt{\quad}$

X = 72 inches

$Z = \frac{\quad}{\quad}$

$= \frac{\quad}{\sqrt{\quad}} = \frac{\quad}{\quad} =$ 1.15023

Table value of 1.15 = 0.3749

Area above 1.15 (above 6 feet) = 0.5 – 0.3749

= 0.125

∴ Number of soldiers who have over 6 feet tall out of 1000 = 0.125 x 1000 = 125

Qn: An aptitude test was conducted for selecting officers in 4 bank from 1000 students. The average score is 42 and the Standard Deviation is 24. Assume normal distibtion for scores and find:-

(a) The number of candidates whose score exceed 58.
(b) The number of candidates whose score lie between 30 and 66.

Sol: (a) Given N = 1000

$\mu = 42$

$\sigma = 24$

$X = 58$

$Z = \frac{X - \mu}{\sigma}$

$= \frac{58 - 42}{24} = \frac{16}{24} = 0.667$

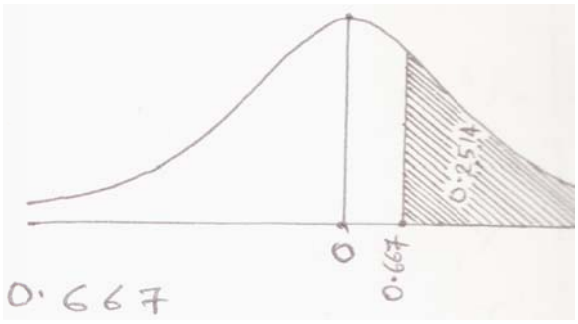


Table value of 0.667 (Area between 0 and 0.667) = 0.2486

$\therefore$ Area above 0.667 = 0.5 – 0.2486

= 0.2514

$\therefore$ Number of students whose score exceed 58 = 0.2514 x 1000

= 251.4

= 251 students

(b) Given N = 1000

$\mu = 42$

$\sigma = 24$

X1 = 30

X2 = 66

$Z = \frac{X - \mu}{\sigma}$

$Z1 = \frac{30 - 42}{24} = \frac{-12}{24} = -0.5$

$Z2 = \frac{66 - 42}{24} = \frac{24}{24} = 1$

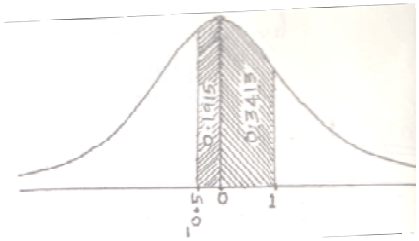


Table value of 0.5 (Area between 0 and -0.5) = 0.1915

Table value of 1 (Area between 0 and +1) = 0.3413

$\therefore$ Area between -0.5 and +1 = 0.1915 + 0.3413

= 0.5328

$\therefore$ Number of students whose score lie between 30 and 66 = 0.5328 x 1000

= 532.8

= 533 students

Fitting of a Normal Distribution

Procedure :

1. Find the mean and standard deviation of the given distribution. (i.e., μ and σ)
2. Take the lower limit of each class.
3. Find Z value for each of the lower limit.
   $Z = \frac{x - \mu}{\sigma}$
4. Find the area for z values from the table. The first and the last values are taken as 0.5.
5. Find the area for each class. Take difference between 2 adjacent values if same signs and take total of adjacent values if opposite signs.
6. Find the expected frequency by multiplying area for each class by N.

Qn: Fit a normal distribution of the following data:

Marks	:	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of students	:	4	22	48	66	40	16	4

Sol:

Computation of mean and standard deviation							
Marks (x)	Mid (m) Point	No. of students (f)	d (m – 35)	d’	fd’	d'^2	fd'^2
10 – 20	15	4	-20	-2	-8	4	16
20 – 30	25	22	-10	-1	-22	1	22
30 – 40	35	48	0	0	0	0	0
40 – 50	45	66	10	1	66	1	66
50 – 60	55	40	20	2	80	4	160
60 – 70	65	16	30	3	48	9	144
70 – 80	75	4	40	4	16	16	64
		N = 200			Σfd’ =180		$\Sigma fd'^2$ =472

$\bar{x}$ = Assumed mean + $\frac{\Sigma f d'}{N}$ + C

= 35 + $\frac{180}{200}$ x 100

= 35 + 9

= 44

$$\sim = \sqrt{\frac{\Sigma f d\prime^2}{N} + \left(\frac{\Sigma f d\prime}{N}\right)^2} \text{ x C}$$

$$= \sqrt{\frac{472}{200} + \left(\frac{180}{200}\right)^2} \text{ x 10}$$

$$= \sqrt{2.36 - 0.9^2} \text{ x 10}$$

$$= \sqrt{2.36 - 0.81} \text{ x 10}$$

$$= \sqrt{1.55} \text{ x 10}$$

$$= 1.245 \text{ x } 10 = \underline{12.45}$$

Computation of expected frequencies				
Lower class limit	$Z = \frac{x-\mu}{\sigma}$	Area under normal curve	Area for each class	Expected frequency (4) x 200
1	2	3	4	5
10	-2.73	0.5000	-	-
20	-1.93	0.4732	0.0268	5
30	-1.12	0.3686	0.1046	21
40	-0.32	0.1255	0.2431	49
50	+0.48	0.1844	0.3099	62
60	+1.29	0.4015	0.2171	43
70	+2.09	0.4817	0.0802	16
80	+2.89	0.5000	0.0183	4
			Total	**200**

CHAPTER - 9

TESTING OF HYPOTHESIS

Statistical Inference:

Statistical inference refers to the process of selecting and using a sample statistic to draw conclusions about the population parameter. Statistical inference deals with two types of problems.

They are:-

1. Testing of Hypothesis
2. Estimation

Hypothesis:

Hypothesis is a statement subject to verification. More precisely, it is a quantitative statement about a population, the validity of which remains to be tested. In other words, hypothesis is an assumption made about a population parameter.

Testing of Hypothesis:

Testing of hypothesis is a process of examining whether the hypothesis formulated by the researcher is valid or not. The main objective of hypothesis testing is whether to accept or reject the hypothesis.

Procedure for Testing of Hypothesis:

The various steps in testing of hypothesis involves the following :-

1. Set Up a Hypothesis:

The first step in testing of hypothesis is to set p a hypothesis about population parameter. Normally, the researcher has to fix two types of hypothesis. They are null hypothesis and alternative hypothesis.

Null Hypothesis:-

Null hypothesis is the original hypothesis. It states that there is no significant difference between the sample and population regarding a particular matter under consideration. The word "null" means 'invalid' of 'void' or 'amounting to nothing'. Null hypothesis is denoted by H_o. For example, suppose we want to test whether a medicine is effective in curing cancer. Hence, the null hypothesis will be stated as follows:-

H_0: The medicine is not effective in curing cancer (i.e., there is no significant difference between the given medicine and other medicines in curing cancer disease.)

Alternative Hypothesis:-

Any hypothesis other than null hypothesis is called alternative hypothesis. When a null hypothesis is rejected, we accept the other hypothesis, known as alternative hypothesis. Alternative hypothesis is denoted by H_1. In the above example, the alternative hypothesis may be stated as follows:-

H_1: The medicine is effective in curing cancer. (i.e., there is significant difference between the given medicine and other medicines in curing cancer disease.)

2. **Set up a suitable level of significance**:

After setting up the hypothesis, the researcher has to set up a suitable level of significance. The level of significance is the probability with which we may reject a null hypothesis when it is true. For example, if level of significance is 5%, it means that in the long run, the researcher is rejecting true null hypothesis 5 times out of every 100 times. Level of significance is denoted by α (alpha).

α = Probability of rejecting H_0 when it is true.

Generally, the level of significance is fixed at 1% or 5%.

3. **Decide a test criterion**:

The third step in testing of hypothesis is to select an appropriate test criterion. Commonly used tests are z-test, t-test, X^2 – test, F-test, etc.

4. **Calculation of test statistic**:

The next step is to calculate the value of the test statistic using appropriate formula. The general fromfor computing the value of test statistic is:-

$$\text{Value of Test statistic} = \frac{\text{Difference}}{\text{Standard Error}}$$

5. **Making Decision**:

Finally, we may draw conclusions and take decisions. The decision may be either to accept or reject the null hypothesis.

If the calculated value is more than the table value, we reject the null hypothesis and accept the alternative hypothesis.

If the calculated value is less than the table value, we accept the null hypothesis.

Sampling Distribution

The distribution of all possible values which can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population is called Sampling distribution of that statistic.

Standard Error (S.E)

Standard Error is the standard deviation of the sampling distribution of a statistic. Standard error plays a very important role in the large sample theory. The following are the important uses of standard errors:-

1. Standard Error is used for testing a given hypothesis

2. S.E. gives an idea about the reliability of a sample, because the reciprocal of S.E. is a measure of reliability of the sample.

3. S.E. can be used to determine the confidence limits within which the population parameters are expected to lie.

Test Statistic

The decision to accept or to reject a null hypothesis is made on the basis of a statistic computed from the sample. Such a statistic is called the test statistic. There are different types of test statistics. All these test statistics can be classified into two groups. They are

a. Parametric Tests

b. Non-Parametric Tests

PARAMETRIC TESTS

The statistical tests based on the assumption that population or population parameter is normally distributed are called parametric tests. The important parametric tests are:-

1. z-test
2. t-test
3. f-test

Z-test:

Z-test is applied when the test statistic follows normal distribution. It was developed by Prof.R.A.Fisher. The following are the important uses of z-test:-

1. To test the population mean when the sample is large or when the population standard deviation is known.
2. To test the equality of two sample means when the samples are large or when the population standard deviation is known.
3. To test the population proportion.
4. To test the equality of two sample proportions.
5. To test the population standard deviation when the sample is large.
6. To test the equality of two sample standard deviations when the samples are large or when population standard deviations are known.
7. To test the equality of correlation coefficients.

Z-test is used in testing of hypothesis on the basis of some assumptions. The important assumptions in z-test are:-

1. Sampling distribution of test statistic is normal.

2. Sample statistics are dose the population parameter and therefore, for finding standard error, sample statistics are used in place where population parameters are to be used.

T-test:

t-distribution was originated by W.S.Gosset in the early 1900.

t-test is applied when the test statistic follows t-distribution.

Uses of t-test are:-

1. To test the population mean when the sample is small and the population s.D.is unknown.
2. To test the equality of two sample means when the samples are small and population S.D. is unknown.
3. To test the difference in values of two dependent samples.
4. To test the significance of correlation coefficients.

The following are the important assumptions in t-test:-

1. The population from which the sample drawn is normal.
2. The sample observations are independent.
3. The population S.D.is known.
4. When the equality of two population means is tested, the samples are assumed to be independent and the population variance are assumed to be equal and unknown.

F-test:

F-test is used to determined whether two independent estimates of population variance significantly differ or to establish both have come from the same population. For carrying out the test of significance, we calculate a ration, called F-ratio. F-test is named in honour of the great statistician R.A.Fisher. It is also called Variance Ration Test.

F-ratio is defined as follows:-

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

$$\text{where } S_1^2 = \frac{(X_1 - \bar{X}_1)^2}{n_1 - 1}$$

$$S_2^2 = \frac{(X_2 - \bar{X}_2)^2}{n_2 - 1}$$

While calculating F-ratio, the numerator is the greater variance and denominator is the smaller variance. So,

$$F = \frac{\text{Greater Variance}}{\text{Smaller Variance}}$$

Uses of F-distribution:-

1. To test the equality of variances of two populations.
2. To test the equality of means of three or more populations.
3. To test the linearity of regression

Assumptions of F-distribution:-

1. The values in each group are normally distributed.
2. The variance within each group should be equal for all groups. ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 \ldots$)
3. The error (Variation of each value around its own group mean) should be independent for each value.

TYPES OF ERRORS IN TESTING OF HYPOTHESIS:

In any test of hypothesis, the decision is to accept or reject a null hypothesis. The four possibilities of the decision are:-

1. Accepting a null hypothesis when it is true.
2. Rejecting a null hypothesis when it is false.
3. Rejecting a null hypothesis when it is true.
4. Accepting a null hypothesis when it is false.

Out of the above 4 possibilities, 1 and 2 are correct, while 3 and 4 are errors. The error included in the above 3rd possibility is called type I error and that in the 4th possibility is called type II error.

Type I Error

The error committed by rejecting a null hypothesis when it is true, is called Type I error. The probability of committing Type I error is denoted by α (alpha).

α = Prob. (Type I error)

= Prob. (Rejecting H_0 when it is true)

Type II Error

The error committed by accepting a null hypothesis when it is false is called Type II error. The probability of committing Type II error is denoted by β (beta).

β = Prob. (Type II error)

β = Prob. (Accepting H_0 when it is false)

Small and Large samples

The size of sample is 30 or less than 30, the sample is called small sample.

When the size of sample exceeds 30, the sample is called large sample.

Degree of freedom

Degree of freedom is defined as the number of independent observations which is obtained by subtracting the number of constraints from the total number of observations.

Degree of freedom = Total number of observations – Number of constraints.

Rejection region and Acceptance region

The entire area under a normal curve may be divided into two parts. They are rejection region and acceptance region.

Rejection Region:

Rejection region is the area which corresponds to the predetermined level of significance. If the calculated value of the test statistic falls in the rejection region, we reject the null hypothesis. Rejection region is also called critical region. It is denoted by α (alpha).

Acceptance Region:

Acceptance region is the area which corresponds to $1 - \alpha$.

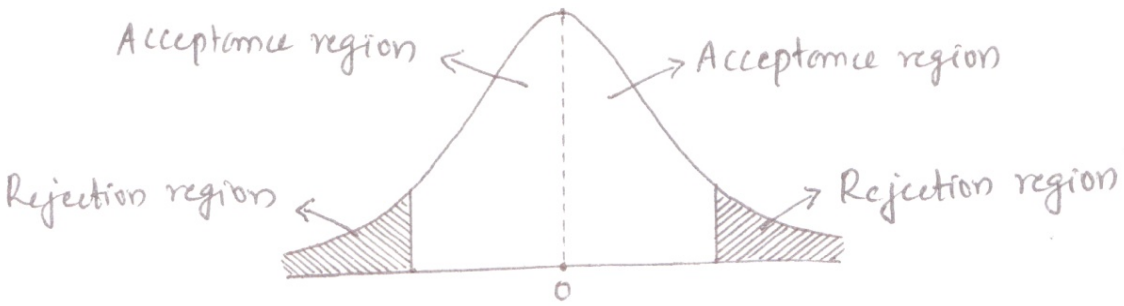
Acceptance region = 1 – rejection region

= $1 - \alpha$.

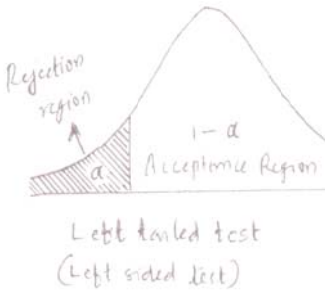
If the calculated value of the test statistic falls in the acceptance region, we accept the null hypothesis.

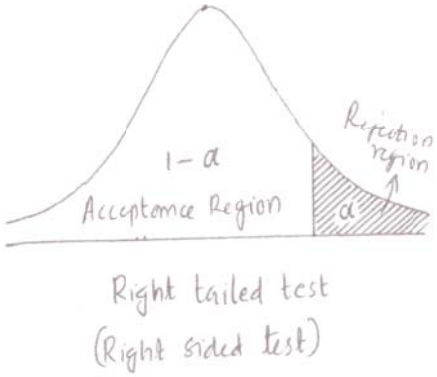
TWO TAILED AND ONE TAILED TESTS:

A two tailed test is one in which we reject the null hypothesis if the computed value of the test statistic is significantly greater or lower than the critical value (table value) of the test statistic. Thus, in two tailed test the critical region is represented by both tails of the normal curve. If we are testing hypothesis at 5 % level of significance, the size of the acceptance region is 0.95 and the size of the rejection region is 0.05 on both sides together. (i.e. 0.025 on left side and 0.025 on right side of the curve).



In one tailed test, the rejection region will be located in only one tail of the normal curve which may be either left or right, depending on the alternative hypothesis. Suppose if the level of significance is 5%, then in case of one tailed test the size of the rejection region is 0.05 either falling in the left side only or in the right side only.





TESTING OF GIVEN POPULATION MEAN

This test is used to test whether the given population mean is true or not. In other words, this test is used to check whether the difference between sample mean and population mean is significant or it is only due to sampling fluctuations. Here we can apply z-test of t-test.

Procedure:

1. Set the null hypothesis that there is no significant difference between sample mean and population mean.

   $H_0 : \mu = \mu_0$

   $H_1 : \mu \neq \mu_0$

2. Decide the test criterion.
   - If sample is large, apply Z-test
   - If sample is small, but population standard deviation is known, apply z test
   - If sample is small and population standard deviation is unknown, apply t-test.
3. Apply the formula

   Z or $t = \frac{\text{Difference}}{SE}$

   $= \frac{\bar{x} - \mu}{SE}$ where $\bar{x}$ Sample mean

   μ = Population mean

   SE = Standard Error

SE is computed as follows:-

$SE = \frac{\sigma}{\sqrt{n}}$ (when population standard deviation is known; sample may be large or small).

SE = $\frac{S}{\sqrt{n}}$ (when population standard deviation is unknown & sample is large).

SE = $\frac{S}{\sqrt{n-1}}$ (when population standard deviation is known and sample is small).

where σ= Population S.D.; S = Sample S.D.; n = Sample size;

4. Fix the degree of freedom.

   For Z-test: Infinity

   For t-test: n-1

5. Obtain the table value at level of significance for degree of freedom.
6. Decide whether to accept or reject the H_0. If calculated value is less than the table value, we accept the H_0 otherwise reject it.

Qn: A sample of 900 items is taken from a population with S.D.15. The mean of the sample is 25. Test whether the sample has come from a population with mean 26.8.

Sol: $H_0 : \mu = 26.8$

$H_1 : \mu \neq 26.8$

Since sample is large apply z-test.

$$Z = \frac{\bar{X} - \mu}{SE} = \frac{\text{Difference}}{SE}$$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{900}} = \frac{15}{30} = 0.5$$

$$\therefore Z = \frac{26.8-25}{0.5}$$

$$= \frac{1.8}{0.5} = 3.6$$

The value of z at 5% level of significance for infinity d.f. is 1.96. As the calculated value is more than the table value , we reject the H_0. There is significant difference sample mean and population mean.

∴ We conclude that the sample has not come from the population with mean 26.8.

Qn. The mean life of 100 bulbs produced by a company is computed to be 1570 hours with S.D. of 120 hours. The company claims that the average life of bulbs produced by the company is 1600 hours. Using 5% level of significance, is the claim acceptable?

Sol: $H_0 : \mu = 1600$

$H_1 : \mu \neq 1600$

Since sample is large apply z-test.

$$Z = \frac{\text{Difference between } \bar{X} \text{ and } \mu}{SE}$$

$$SE = \frac{S}{\sqrt{n}} = \frac{120}{\sqrt{100}} = \frac{120}{10} = \underline{\underline{12}}$$

$\therefore z = \frac{1600-1570}{12}$

$= \frac{30}{12} = \underline{\underline{2.5}}$

Table value at 5% level of significance and infinity d.f. is 1.96. As the calculated value is greater than the table value, we reject the H_0. There is significant difference between mean life of sample and mean life of population.

∴ Company's claim is not acceptable

Qn. The price of shares of a company on the different days in a month were found to be 66,65,69,70,69,71,70,63,64 and 68. Discuss whether mean price of shares in the month is 65.

Sol. $H_0 : \mu = 65$

$H_1 : \mu \neq 65$

Since small sample, apply t-test.

$t = \frac{\text{Difference between } \bar{X} \text{ and } \mu}{SE}$

$\bar{X} = \frac{\sum X}{N} = \frac{66+65+69+70+69+71+70+63+64+68}{10}$

$= \frac{675}{10} = \underline{\underline{67.5}}$

Computation of Standard deviation		
X	$X-\bar{X}$	$(X-\bar{X})$
66	-1.5	2.25
65	-2.5	6.25
69	1.5	2.25
70	2.5	6.25
69	1.5	2.25
71	3.5	12.25
70	2.5	6.25
63	-4.5	20.25
64	-3.5	12.25
68	0.5	0.25
		70.50

Standard deviation (S) = $\sqrt{\frac{\sum(X-\bar{X})^2}{N}}$

$= \sqrt{\frac{70.50}{10}} = \sqrt{7.05}$

$= \underline{\underline{2.655}}$

S.E $= \frac{S}{\sqrt{n-1}}$

$= \frac{2.655}{\sqrt{10-1}} \quad = \frac{2.655}{\sqrt{9}} \quad = \frac{2.655}{3}$

$= \underline{\underline{0.885}}$

$\therefore t = \frac{67.5-65}{0.885} = \frac{2.5}{0.885}$

$= \underline{\underline{2.8249}}$

Table value at 5% level of significance and 9 d.f. = 2.262. Calculated value is more than table value.

∴ We reject the null hypothesis.

We conclude that the mean price of share in the month is not 65/-.

Qn. The mean height obtained from a random sample of 36 children is 30 inches. The standard deviation of the distribution of height of the population is known to be 1.5 inches. Test the statement that the mean height of the population is 33 inches at 5% level of significance. Also set up 99% confidence limits of the mean height of the population.

Sol. $H_0 : \mu = 33$

$H_1 : \mu \neq 33$

Since small sample is large, apply z-test.

$z = \frac{\text{Difference between } \bar{X} \text{ and } \mu}{SE}$

$\bar{X} = 30$

$\mu = 33$

$SE = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{36}} = \frac{1.5}{6} = \underline{\underline{0.25}}$

$\therefore z = \frac{33-30}{0.25} = \frac{3}{0.25} = \underline{\underline{12}}$

Table value at 5% level of significance & infinity d.f. = 1.95.

Calculated value is greater than table value.

∴ We reject the H_0

∴ We conclude that this mean height of population is not 33 inches.

(b) 99% confidence level = 1% significance level.

Table value of Z at 1% level of significance and infinity d.f. } = 2.576

Limits of Population $= \bar{X} \pm 2.576$ SE

$= 30 \pm (2.576 \times 0.25)$

$= 30 \pm 0.644$

= 29.356 and 30.644

Qn. An investigation of a sample of 64 BBA students indicated that the mean time spend on preparing for the examination was 48 months and the S.D. was 15 months. What is the average time spent by all BBA students before they complete their examinations.

Sol. Here, the students are asked to compute the population mean. In other words, the confidence limits of population mean.

Table value of z at 5% level of significance and infinity d.f. } = 1.96

Confidence limits of μ $= \mu \pm 1.96$ SE

S.E. $= \frac{S}{\sqrt{n}} = \frac{15}{\sqrt{64}} = \frac{15}{8} = 1.875$

∴ Confidence Limits of μ $= 48 \pm 1.96 \times 1.875$

$= 48 \pm 3.675$

= 44.325 and 51.675

∴ Average time spent by all BBA students before competing their examinations is between 44.325 months and 51.675 months.

Qn. A typist claims that he can take dictations at the rate of more than 120 words per minute. Of the 12 tests given to him, he could perform an average of 135 words with a S.D. of 40. Is his claim valid. (use 1% level of significance).

$H_0 : \mu = 120$

$H_1 : \mu > 120$

Since small sample is large, apply z-test.

$t = \frac{\text{Difference}}{SE} = \frac{135-120}{SE}$

SE $= \frac{S}{\sqrt{n-1}} = \frac{40}{\sqrt{12-1}} = \frac{40}{\sqrt{11}} = \frac{40}{3.317} = \underline{12.06}$

$\therefore t = \frac{135-120}{12.06} = \frac{15}{12.06} = \underline{1.24}$

Table value of 't' at 1% level of significance and 11 d.f. } = 2.718

Calculated value is less than table value.

∴ We accept the null hypothesis i.e, μ = 120

∴ We conclude that his claim of taking dictation at the rate of more than 120 words per minute is not valid.

Qn. A factory was producing electric bulbs of average length of 2000 hours. A new manufacturing process was introduced with the hope of increasing the length of the life of bulbs. A sample of 25 bulbs produced by the new process were examined and the average length of life was found to be 2200 hours. Examine whether the average length of bulbs was increased assuming the length of lives of bulbs follow normal distribution with $\sigma = (\alpha\ 0.05)$.

Sol. $H_0 : \mu = 2000$

$H_1 : \mu > 2000$

Since sample is small, apply test.

$z = \frac{\text{Difference}}{SE}$

SE $= \frac{\sigma}{\sqrt{25}} = \frac{300}{5} = \underline{60}.$

$t = \frac{2200-2000}{60} = \frac{200}{60} = \underline{3.33}$

Table value of 't' at 5% significance level and 24 d.f. = 1.711

Calculated value is greater than the table value.

∴ We reject the null hypothesis and accept alternative hypothesis. So we conclude that the new manufacturing process has increased the life of bulbs, i.e, μ = 200.

TESTING OF EQUALITY OF TWO SAMPLE MEANS

This test is used to test whether there is significant difference between two sample means. It there is no significant difference, we can consider the samples are drawn from the same population.

Procedure:

1. Set up null hypothesis that there is no significant difference between the tow means.

$H_0 : \mu 1 = \mu 2$

$H_1 : \mu 1 \neq \mu 2$

2. Decide the test criterion:
   - If sample is large, apply z – test
   - If sample is small, but population S.D. is known, apply z-test.
   - If sample is small and population S.D. is unknown, apply t-test.
3. Apply the formula:

$$Z \text{ or } t = \frac{\text{Difference between means}}{SE} = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

SE is computed as follows:

- It population S.D. are known and equal, S.E. = $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$
- It population S.D. are known but different, S.E. = $\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$
- It population S.D. are unknown and samples are large, then assuming population S.D. are different,

  S.E. = $\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}$
- It population S.D. are unknown and samples are small, then assuming population S.D. are equal,

  S.E. = $\sqrt{\frac{n_1 {S_1}^2 + n_2 {S_2}^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

4. Fix the degree of freedom:

   For Z-test : Infinity

   For t-test: $n_1 + n_2 - 2$

5. Obtain the table value.

6. Decide whether to accept or reject the H_0.

Qn: Fifty children were given special diet for a certain period and control group of 50 other children were given normal diet. Their average gain in weight were found to be 7.2 kgs and 5.7 kgs respectively and the common S.D. for gain in weight was 2 Kgs. Assuming normality of the distributions whould you conclude that the diet really promoted weight?

Sol: H_0 : Special diet does not promote weight; $\mu_1 = \mu_2$

H_1 : Special diet promotes weight; $\mu_1 > \mu_2$

Since samples are large, apply z- test.

$$Z = \frac{\text{Difference between Samples}}{SE} = \frac{7.2 - 5.7}{SE}$$

$$SE = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$= \sqrt{\frac{2^2}{50} + \frac{2^2}{50}} = \sqrt{\frac{4}{50} + \frac{4}{50}} = \sqrt{\frac{8}{50}}$$

$$= \sqrt{0.16} = \underline{0.4}$$

$$\therefore Z = \frac{7.2-5.7}{0.4} = \frac{1.5}{0.4} = \underline{\underline{3.75}}$$

Table value of Z at 5% level of significance and infinity degree of freedom is 1.645.

Calculated value is greater than the table value.

∴ We reject the null hypothesis; and accept H_1. i.e. $\mu_1 > \mu_2$.

So we conclude that the special diet really promotes weight.

Qn. The mean weight of a sample of 80 boys of class X was found to be 65 kg with a S.D. of 7 Kg. Another sample of 85 boys of class X shows a mean weight of 69 Kg. with a S.D. of 5 Kg. Can the two samples be considered as drawn from the same population whose S.D. is 6 Kg. Test at 5% level of significance.

Sol. Here population S.D. are given equal.

H_0 : There is no significant difference in mean weight of 2 samples; $\mu_1 = \mu_2$

H_1 : There is significant different in mean weight of two samples; $\mu_1 \neq \mu_2$

Samples are large, apply z- test.

$$Z = \frac{\text{Difference between means}}{SE} = \frac{69 - 65}{SE}$$

$$SE = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$= \sqrt{\frac{6^2}{80} + \frac{6^2}{85}} = \sqrt{\frac{36}{80} + \frac{36}{80}} = \sqrt{0.45 + 0.424}$$

$$= \sqrt{0.874} = \underline{\underline{0.9349.}}$$

$$\therefore Z = \frac{69-65}{0.9349} = \frac{4}{0.9349} = \underline{\underline{4.2785.}}$$

Table value of Z at 5% level of significance and infinity d.f. = 1.96.

Calculated value is greater than the table value.

∴ We reject the H_0 and accept H_1.

So, we conclude that the samples are not drawn from the population having the S.D. of 6 Kg.

Qn. Elective bulbs manufactured by X Ltd. and Y Ltd. gave the following results.

	X Ltd.	**Y Ltd.**
No. of bulbs used	100	100
Mean life in hours	1300	1248
Standard deviation	82	93

Using S.E. of the difference between mean, state whether there is any significant difference in the life of the two makes.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$$Z = \frac{\text{Difference between means}}{SE}$$

$$SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$= \sqrt{\frac{82^2}{100} + \frac{93^2}{100}} = \sqrt{\frac{6724}{100} + \frac{8649}{100}} = \sqrt{67.24 + 86.49}$$

$$= \sqrt{153.73} = \underline{12.399.}$$

$$\therefore Z = \frac{1300-1248}{12.399} = \frac{52}{12.399} = \underline{4.194.}$$

Table value of Z = 1.96.

Calculated value is greater than the table value.

∴ We reject the H_0

So we conclude that there is significant difference in the average life of bulbs of 2 makes.

Qn. The average number of articles manufactured by two machines per day and 200 and 250 with S.D. 20 and 25 respectively on the basis of 25 days' production. Can you regard both the machines are equally efficient at 1% level of significance?.

Sol. $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$$Z = \frac{\text{Difference in means}}{SE}$$

Here population S.D. are not known and samples are small.

$$\text{S.E.} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{(25 \times 20^2) + (25 \times 25^2)}{25+25-2}\left(\frac{1}{25}+\frac{1}{25}\right)}$$

$$= \sqrt{\frac{(25 \times 400) + (25 \times 625)}{48}\left(\frac{2}{25}\right)}$$

$$= \sqrt{\frac{25{,}625}{48} \times 0.08}$$

$$= \sqrt{42.7083}$$

$$= \underline{6.5352}.$$

$$\therefore t = \frac{250-200}{6.5352} = \frac{50}{6.5352} = \underline{7.651}$$

Degree of freedom $= n_1 + n_2 - 2$

$= 25 + 25 - 2 = 48$

Table value of 't' at 1% level of significance and infinity d.f. = 2.576.

Calculated value is greater than the table value.

$\therefore$ We reject the null hypothesis.

i.e, $\mu_1 \neq \mu_2$

So we conclude that the machines are not equally efficient.

Qn: Given below are the gains in weights of dogs on two diets, X and Y

Diet X:	15	22	20	22	18	14	22			
Diet Y :	14	24	12	20	32	21	30	20	22	25

Test at 5% level, whether the two diets differ significantly with regard to increase in weight.

Sol: H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Since samples are small, apply t - teat

$$t = \frac{\text{Difference in mean}}{SE}$$

As mean and S.D. are not given, first we have to find out the same of the 2 groups.

Computation of mean and S.D. of Diet X (i.e.,$\bar{X}_1$ and S_1)		
X	**d (X-20)**	**d^2**
15	-5	25
22	2	4
20	0	0
22	2	4
18	-2	4
14	-6	36
22	2	4
	$\Sigma d = -7$	$\Sigma d^2 = 77$

$$\bar{X}_1 = A + \frac{\Sigma d}{N_2}$$

$$= 20 + \frac{-7}{7} = 20 + -7 = 19$$

$$S_1 = \sqrt{\frac{\Sigma d^2}{N_1} - \left(\frac{\Sigma d}{N_1}\right)^2}$$

$$= \sqrt{\frac{77}{7} - \left(\frac{7}{7}\right)^2} = \sqrt{11 - 1}$$

$$= \sqrt{10} = \underline{\underline{3.1623}}$$

Computation of mean and S.D. of Diet X (i.e., $\bar{X}_2$ and S_2)		
X	**d (X-20)**	**d^2**
14	-6	36
24	4	8
12	-8	64
20	0	0
32	12	144
21	1	1
30	10	100
20	0	0
22	2	4
25	5	25
	$\Sigma d = 20$	$\Sigma d^2 = 390$

$$\bar{X}_2 = A + \frac{\sum d}{N_1}$$

$$= 20 + \frac{20}{10} = 20 + 2 = 22$$

$$S_2 = \sqrt{\frac{\Sigma d^2}{N_2} - \left(\frac{\Sigma d}{N_2}\right)^2}$$

$$= \sqrt{\frac{390}{10} - \left(\frac{20}{10}\right)^2} = \sqrt{39 - 4}$$

$$= \sqrt{35} = 5.9161$$

$$\text{S.E.} = \sqrt{\frac{n_1 {S_1}^2 + n_2 {S_2}^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{(7 \times 3.1623^2) + (10 \times 5.9161^2)}{7 + 10 - 2}\left(\frac{1}{7} + \frac{1}{10}\right)}$$

$$= \sqrt{\frac{(7\times10)+(10\times35)}{15}\,(0.1429+0.1)}$$

$$= \sqrt{\frac{420}{15}(0.2429)} = \sqrt{28 \times 0.2429}$$

$$= \sqrt{6.8012} = \underline{\underline{2.6079}}$$

$$\therefore t = \frac{\text{Difference in mean}}{\text{SE}}$$

$$= \frac{22-19}{2.6079}$$

$$= \frac{3}{2.6079}$$

$$= \underline{\underline{1.1504}}$$

Table value of 't' at 5% level of significance and 15 d.f. (7 + 10-2) } 2.731

Calculated value is less than table value.

∴ We accept the null hypothesis i.e, $\mu_1 = \mu_2$

So, we can conclude that there is no significant difference between two diets.

TESTING OF EQUALITY OF TWO SAMPLE STANDARD DEVIATIONS

This test is used to test whether there is any significant difference between the standard between the standard deviation of two samples.

Procedure:

1. Set the null hypothesis that there is no significant difference between two standard deviations.

   H_0: $\boldsymbol{\sigma_1 = \sigma_2}$

   H_0: $\boldsymbol{\sigma_1 \neq \sigma_2}$
2. Decide the test criterion:
   If sample is large, apply Z – test
   If sample is sample, apply F – test
3. Apply the formula:
   If Z test:

$$Z = \frac{\text{Difference in S.D}}{\text{S.E}} = \frac{S_1 - S_2}{\text{S.E}}$$

$$SE = \sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}$$ (When population S.D. are known)

$$SE = \sqrt{\frac{{S_1}^2}{2n_1} + \frac{{S_2}^2}{2n_2}}$$ (When population S.D. are not known)

If F – test:

$$F = \frac{\frac{n_1 {S_1}^2}{n_1 - 1}}{\frac{n_2 {S_2}^2}{n_2 - 1}}$$ (Larger value must be numerator and smaller must be denominator)

4. Fix the degree of freedom

   For Z – test: Infinity

   For F – test: $(n_1 - 1, n_2 - 1)$

5. Obtain the table value.

6. Decide whether to accept or reject the null hypothesis.

Qn: A sample of 60 items has S.D of 5 and another sample of 80 items has S.D of 4.5. Can you assert that the two samples belong to the same population?

Sol: H_0: $\boldsymbol{\sigma_1 = \sigma_2}$

H_0: $\boldsymbol{\sigma_1 \neq \sigma_2}$

Since samples are large, apply Z - test.

$$Z = \frac{\text{Difference in S.D}}{\text{S.E}}$$

$$SE = \sqrt{\frac{{S_1}^2}{2n_1} + \frac{{S_2}^2}{2n_2}}$$

$$= \sqrt{\frac{5^2}{2\times60} + \frac{4.5^2}{2\times80}} \quad = \sqrt{\frac{25}{120} + \frac{20.5}{160}}$$

$$= \sqrt{0.2083 + 0.1266} \quad = \sqrt{0.3349}$$

$$= 0.5787$$

$$\therefore Z = \frac{5-4.5}{0.5787} = \frac{0.5}{0.5787} = \underline{0.8640}$$

Table value of Z at 5% level of significance and infinity degree of freedom } = 1.96

Calculated value is less than the table value.

∴ We accept the null hypothesis.

So, we conclude that there is no significant difference between standard deviation and the two samples belong to the same population.

Qn: The S.D. of two samples of sizes 10 and 14 from two normal populations are 3.5 and 3 respectively. Examine whether the S.D of the populations are equal.

Sol: H_0: $\boldsymbol{\sigma_1 = \sigma_2}$

H_0: $\boldsymbol{\sigma_1 \neq \sigma_2}$

Since samples are small, apply F-test.

$$F = \frac{\frac{n_1 S_1^2}{n_1 - 1}}{\frac{n_2 S_2^2}{n_2 - 1}}$$

$$\frac{n_1 S_1^2}{n_1 - 1} = \frac{10 \times 3.5^2}{10-1} = \frac{10 \times 12.25}{9} = \frac{122.5}{9} = \underline{13.61}$$

$$\frac{n_2 S_2^2}{n_2 - 1} = \frac{14 \times 3^2}{14-1} = \frac{14 \times 9}{13} = \frac{1126}{13} = 9.69$$

$$\therefore F = \frac{13.61}{9.69} = \underline{1.405}$$

Degree of freedom is $(n_1 - 1, n_2 - 1)$ = (10-1, 14-1)

= (9, 13)

∴ Table value of F at 5% level of significance and (9, 13) d.f. = 2.72

Calculated value of F is smaller than the table value.

∴ We accept the H_0 i.e., $\boldsymbol{\sigma_1 = \sigma_2}$

So we conclude that the S.D. of the populations are equal.

Qn: Two random sample were drawn from two normal populations and their values are :-

A:	66	67	75	76	82	84	88	90	92		
B:	64	66	74	78	82	85	87	92	93	95	97

Examine whether the standard deviations of the population are equal.

Sol: Here, first we find out the standard deviations

Computation of S.D. of 2 samples					
Sample A			Sample B		
X	d (X-75)	d^2	X	d (X-82)	d^2
66	-9	81	64	-18	324
67	-8	84	66	-16	256
75	0	0	74	-8	64
76	1	1	78	-4	16
82	7	49	82	0	0
84	9	81	85	3	9
88	13	169	87	5	25
90	15	225	92	10	100
92	17	289	93	11	121
	Σd = 45	Σd^2= 959	95	13	169
			97	15	225
				Σd=11	Σd^2 = 1309

$$\text{S.D} = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

$$\therefore \text{SD of sample A} = \sqrt{\frac{959}{9} - \left(\frac{45}{9}\right)^2} = \sqrt{106.56 - 25}$$

$$= \sqrt{81.56} = 9.03$$

$$\text{SD of sample B} = \sqrt{\frac{1309}{11} - \left(\frac{11}{11}\right)^2} = \sqrt{119 - 1}$$

$$= \sqrt{118} = 10.86$$

H_0: $\boldsymbol{\sigma_1 = \sigma_2}$

H_0: $\boldsymbol{\sigma_1 \neq \sigma_2}$

Since sample are small, apply F-test

$$F = \frac{\frac{n_1 {S_1}^2}{n_1 - 1}}{\frac{n_2 {S_2}^2}{n_2 - 1}}$$

$$\frac{n_1 {S_1}^2}{n_1 - 1} = \frac{9 \times (9.03)^2}{9-1} = \frac{9 \times 81.54}{8} = \frac{733.86}{8} = \underline{\underline{91.7325}}$$

$$\frac{n_2 {S_2}^2}{n_2 - 1} = \frac{11 \times 10.86^2}{11-1} = \frac{11 \times 117.94}{10} = \frac{1297.34}{10} = \underline{\underline{129.734}}$$

$$\therefore F = \frac{129.734}{91.7325} = \underline{\underline{1.414}}$$

Table value of F at 5% level of significance of (10, 8) d.f = 3.34

Table value is greater than the calculated value.

$\therefore$ We accept the H_0

So, we conclude that the S.D. of the populations are equal.

CHAPTER 10

NON-PARAMETRIC TESTS

A non-parametric test is a test which is not concerned with testing of parameters. Non-parametric tests do not make any assumption regarding the form of the population. Therefore, non-parametric tests are also called distribution free tests.

Following are the important non-parametric tests:-

1. Chi-square test ($\chi^2 - test$)
2. Sign test
3. Signed rank test (Wilcoxon matched pairs test)
4. Rank sum test (Mann-whitney U-test and Kruskal-Wallis H test)
5. Run test
6. Kolmogrov-Smirnor Test (K-S-test)

CHI-SQUARE TEST ($\chi^2 - test$)

The value of chi-square describes the magnitude of difference between observed frequencies and expected frequencies under certain assumptions. χ^2 value (χ^2 quantity) ranges from zero to infinity. It is zero when the expected frequencies and observed frequencies completely coincide. So greater the value of χ^2, greater is the discrepancy between observed and expected frequencies.

χ^2-test is a statistical test which tests the significance of difference between observed frequencies and corresponding theoretical frequencies of a distribution without any assumption about the distribution of the population. This is one of the simplest and most widely used non-parametric test in statistical work. This test was developed by Prof. Karl Pearson in 1990.

Uses of χ^2 - test

The uses of chi-square test are:-

1. Useful for the test of goodness of fit:- χ^2 - test can be used to test whether there is goodness of fit between the observed frequencies and expected frequencies.
2. Useful for the test of independence of attributes:- χ^2 test can be used to test whether two attributes are associated or not.
3. Useful for the test of homogeneity:- χ^2 -test is very useful t5o test whether two attributes are homogeneous or not.
4. Useful for testing given population variance:- χ^2-test can be used for testing whether the given population variance is acceptable on the basis of samples drawn from that population.

χ^2 -test as a test of goodness of fit:

As a non-parametric test, χ^2-test is mainly used to test the goodness of fit between the observed frequencies and expected frequencies.

Procedure:-

1. Set up mull hypothesis that there is goodness of fit between observed and expected frequencies.
2. Find the χ^2 value using the following formula:-

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where O = Observed frequencies

E = Expected frequencies

3. Compute the degree of freedom.

d. f. = n – r – 1

Where 'r' is the number of independent constraints to be satisfied by the frequencies

4. Obtain the table value corresponding to the lord of significance and degrees of freedom.
5. Decide whether to accept or reject the null hypothesis. If the calculated value is less than the table value, we accept the null hypothesis and conclude that there is goodness of fit. If the calculated value is more than the table value we reject the null hypothesis and conclude that there is no goodness of fit.

Qn:- A sample analysis of examination result of 200 students were made. It was found that 46 students had failed, 68 secured IIIrd class, 62 IInd class and the rest were placed in the Ist class. Are these figures commensurate with the general examination results which is in the ratio of 2 : 3: 3: 2 for various categories respectively?

Sol: H_0: The figures commensurate with the general examination results.

H_1: The figures do not commensurate with the general examination results.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Computation of χ^2 value:				
O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
46	200 x $\frac{2}{10}$ = 40	6	36	0.9000
68	200 x $\frac{2}{10}$ = 40	8	64	1.0667
62	200 x $\frac{2}{10}$ = 40	2	4	0.0667
24	200 x $\frac{2}{10}$ = 40	-16	256	6.4000
			$\sum\frac{(O-E)^2}{E}$	= 8.4334

χ^2 = 8. 4334

The table value at 5% level of significance and degree of freedom at 3. (df = n – r- 1 =4 – 0 – 1 = 3) } = 7. 815

The calculated value is more than the table value.

∴ we reject the H_0

∴ we conclude that the analytical figures do not commensurate with the general examination result. In other words, there is no goodness of fit between the observed and expected frequencies.

Qn: Test whether the accidents occur uniformity over week days on the basis of the following information:-

Days of the week:	Sun	Mon	Tue	Wed	Thu	Fri	Sat
No. of accidents:	11	13	14	13	15	14	18

Sol: H_0: There is goodness of fit between observed and expected frequencies, i.e., accidents occur uniformly over week days.

H_1: There is no goodness of fit between observed and expected frequencies; i.e., accidents do not accrue uniformly over week days

$$\chi^2 = \sum\frac{(O-E)^2}{E}$$

Computation of χ^2 value:				
O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
11	14	-3	9	0.6429
13	14	-1	1	0.0714
14	14	0	0	0.0000
13	14	-1	1	0.0714
15	14	1	1	0.0714
14	14	0	0	0.0000
18	14	4	16	1.1429
		$\sum\frac{(O-E)^2}{E}$		= 2.0000

The value of χ^2 at 5% level of significance and n – r- 1 = 7 – 0 – 1 = 6 d.f } = 12.592

Calculated value if less than the table value.

∴we accept the null hypothesis. We may conclude that there is goodness of fit between and expected frequencies. i.e., the accidents occur uniformity over week days.

χ^2 – test as a test of independence:

χ^2 – test is used to find out whether one or more attributes are associated or not.

Procedure:-

1. Set up null and alternative hypothesis.

   H_0: Two attributes are independent (i.e., there is no association between the attributes)

   H_1: Two attributes are dependent (i.e., there is an association between the attributes)

2. Find the χ^2 value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

3. Find the degree of freedom

d.f. = (r-1)(c-1)

Where r = Number of rows

c = Number of columns

4. Obtain table value corresponding to the level of significance and degree of freedom.

5. Describe whether to accept or reject the H_0 . If the calculated value is less than the table value, we accept the H_0 and conclude that the attributes are independent. If the H_0 and conclude that the attributes are dependent.

Qn: The following table gives data regarding election to an office:-

Attitude towards election	Economic Status		
	Rich	Poor	Total
Favourable	50	155	205
Non favourable	90	110	200
Total	140	265	405

Is attitude towards election influenced by economic status of workers?

Sol: H_0: The two attributes, election and economic status are independent.

H_1: The attributes, election and economic status are dependent.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Observed frequencies are 50, 90, 155 and 110

Computation of expected frequencies (2 x 2 contingency table)			
Economic Status → / Attitude towards election ↓	Rich	Poor	Total
Favourable	$\frac{140 \times 205}{405} = 71$	$\frac{205 \times 265}{405} = 134$	205
Not favourable	$\frac{140 \times 200}{405} = 69$	$\frac{200 \times 265}{405} = 405$	200
Total	140	265	405

Computation of χ^2 value:

O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
50	71	-21	441	6.21
90	69	21	441	6.39
155	134	21	441	3.29
110	131	-21	441	3.37
			$\sum\frac{(O-E)^2}{E}$	= 19.26

Table value of χ^2 at 5% level of significance for 1 d.f. is 3.841 (d.f = (2-1)(2-1)=1.

Calculated value is greater than the table value. ∴ we reject the H_0.

∴ Election and economic status are not independent (i.e., dependent)

Qn: In a sample study about the tea habit in two towns, following date are observed in a sample of size 100 each:-

Town –A:-

51 persons were male, 31 were tea drinkers and 19 were male tea drinks.

Town – B :-

46 persons were male, 17 were male tea drinkers and 26 were tea drinkers.

Is there any association between sex and tea habits ?
If so, in which town it is greater?

Sol:- H_0: The two attributes, sex and tea habits are independent.
H_1: The two attributes sex and tea habits are dependent.

Town A:-

2 x 2 Contingency table of observed frequency

Sex / Tea habits	Male	Female	Total
Tea Drinkers	19	12	31
Not tea drinkers	32	37	69
Total	51	49	100

Computation of expected frequencies (2 x 2 contingency table)			
Sex → / Tea Habits ↓	**Male**	**Female**	**Total**
Tea Drinkers	$\frac{51 \times 31}{100} = 16$	$\frac{49 \times 31}{100} = 15$	31
Not tea drinkers	$\frac{51 \times 69}{100} = 35$	$\frac{49 \times 69}{100} = 34$	69
Total	51	49	100

Computation of χ^2 value:

O	E	O - E	$(O - E)^2$	$\frac{(O - E)^2}{E}$
19	16	3	9	0.5625
32	35	-3	9	0.2571
12	15	-3	9	0.6000
37	34	3	9	0.2647
			$\sum \frac{(O-E)^2}{E}$	= 1.6843

Degree of freedom = (c-1)(r-1) = (2-1)(2-1) = 1

Table value of χ^2 at 5% level of significance for 1 degree of freedom is 3.84. As the calculated value is less than the table value we accept the null hypothesis in case of Town A. In other words, sex and tea habits are independent (not associated) in Town A.

Town B:-

Contingency table of observed frequencies

Sex → / Tea habits ↓	**Male**	**Female**	**Total**
Tea Drinkers	17	9	26
Not tea drinkers	29	45	74
Total	46	54	100

Computation of expected frequencies (2 x 2 contingency table)			
Sex → / Tea Habits ↓	**Male**	**Female**	**Total**
Tea Drinkers	$\frac{24 \times 46}{100} = 12$	$\frac{54 \times 26}{100} = 14$	26
Not tea drinkers	$\frac{74 \times 46}{100} = 34$	$\frac{54 \times 74}{100} = 40$	74
Total	46	54	100

Computation of χ^2 value:

O	E	O - E	$(O - E)^2$	$\frac{(O - E)^2}{E}$
17	12	5	25	2.083
29	34	-5	25	0.735
9	14	-5	25	1.786
45	40	5	25	0.625
		$\sum\frac{(O-E)^2}{E}$		5.229

Degree of freedom = (2-1)(2-1) = 1

The table value of χ^2 at 5% level of significance for 1 degree of freedom is 3.84. As the calculated value is more than the table value, we reject the H_0 . In other words, attributes sex and tea habits are not independent (i.e., associated) in Town B.

χ^2 – test as a test of homogeneity

χ^2 – test is used to find whether the samples are homogeneous as far as a particular attribute is concerned.

Steps:

1. Set up null and alternative hypotheses:

   H_0: There is homogeneity.

   H_1: There is no homogeneity (heterogeneity)

2. Find the χ^2 value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

3. Find the degree of freedom

d.f. = (r-1)(c-1)

4. Obtain the table value

5. Decide whether to accept or reject the null hypothesis.

Qn: From the adult population of four large cities, random samples were selected and the number of married and unmarried men were recorded:

Cities

	A	**B**	**C**	**D**	**Total**
Married	137	164	152	147	600
Single	32	57	56	35	180
Total	169	221	208	182	780

Is there significant variation among the cities in the tendency of men to marry.

Sol:- H_0: The 4 cities are homogeneous.

H_1: The 4 cities are heterogeneous.

Computation of expected frequencies

Sex → / **Married Status ↓**	**A**	**B**	**C**	**D**	**Total**
Married	$\frac{169 \times 600}{780} = 130$	$\frac{221 \times 600}{780} = 170$	$\frac{208 \times 600}{780} = 160$	$\frac{182 \times 600}{780} = 140$	600
Single	$\frac{169 \times 180}{780} = 39$	$\frac{221 \times 180}{780} = 51$	$\frac{208 \times 180}{780} = 48$	$\frac{182 \times 180}{780} = 42$	180
Total	169	221	208	182	780

Computation of χ^2 value:

O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
137	130	7	49	0.3769
32	39	-7	49	1.2564
164	170	-6	36	0.2118
57	51	6	36	0.7059
152	160	-8	64	0.4000
56	48	8	64	1.3333
147	140	7	49	0.3500
35	42	-7	49	1.1667
$\sum\frac{(O-E)^2}{E}$				5.8010

Degree of freedom = (r-1)(c-1)

= (2-1)(4-1) = 1 x 3 = 3

The table value at 5% for 3 d.f. = 7.82

As the calculated value χ^2 is less than the table value, we accept the H_0 . The cities are homogeneous. So we conclude that there is no significant variation among cities in the tendency of men to marry.

χ^2 – test for Population Variance:

χ^2 – test can be used for testing the given population when the sample is small.

Steps:-

1. Set up null and alternative hypotheses:

   H_0: There is no significant difference between sample variance and population variance.

   H_1: There is significant difference between sample variance and population variance.

2. Find the χ^2 value.

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

Where s^2 = Sample variance

σ^2 = Population variance

3. Find the degree of freedom

d.f. = n-1

4. Obtain the table value

5. Decide whether to accept or reject the null hypothesis

Qn: The standard deviation of a sample of 10 observations from a normal population was formed to be 5. Examine whether this is consistent with the hypothesis that the standard deviation of the population is 5.3.

Sol: H_0: There is no significant difference between sample standard deviation and population standard deviation.

H_1: There is significant difference between sample standard deviation and population standard deviation.

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{10 \times 5^2}{(5.3)^2} = \frac{250}{28.09} = 8.8999$$

Table value of χ^2 at 5% level of significance of 9 d.f. = 16.9 calculated value of χ^2 is less than the table value , we accept in null hypothesis. So we conclude that there is no significant difference between sample variance and population variance.

Limitations of Chi-square tests:-

1. It is not as reliable as a parametric test. Hence it should be used only when parametric tests cannot be used.
2. χ^2 value can not be computed when the given values are proportions or percentages.

WILCOXON MATCHED PAIRS TEST

SIGNED RANK TEST

Signed rank test was developed by Frank Wilcoxon. It is an important non-parametric test. This method is used when we can determine both direction and magnitude of difference between matched values.

Here there are two cases:-

a) When the number of matched pairs are less than or equal to 25.

b) When the number of matched pairs are more than 25.

Case:1

When the number of matched pairs are less than or equal to 25

Procedure:-

1. Set up null hypothesis:
   H_0: There is no significant difference.
   H_1: There is significant difference.
2. Find the difference between each pair of values.
3. Assign ranks to the differences from the smallest to the largest without any regard to sign.
4. Then actual signs of each difference are put to the corresponding ranks.
5. Find the total of positive ranks and negative ranks.
6. Smaller value, as per steps 5 is taken as the calculated value.
7. Obtain the table value of Wilcoxon's T-Table.
8. Decide whether to accept or reject the null hypothesis.

Qn: Given below is 16 pairs of values showing the performance of two machines A and B. Test whether there is difference between the performances. Table value of 'T' at 5% significanterd is 25.

A:	73, 43, 47, 53, 58, 47, 52, 58, 38, 61, 56, 56, 34, 55, 65, 75
B:	51, 41, 43, 41, 47, 32, 24, 58, 43, 53, 52, 57, 44, 57, 40, 68

Sol: H_0: There is no significant difference between the performance of 2 machines.

H_1: There is significant difference the performance of 2 machines.

1	2	3	4	5	
Machine A	Machine B	Difference (3) = (1) – (2)	Rank of Difference (without signs)	Rank with signs	
				+ Sign	- Sign
73	51	22	13	13	
43	41	2	2.5	2.5	
47	43	4	2.5	4.5	
53	41	12	11	11	
58	47	11	10	10	
47	32	15	12	12	
52	24	28	15	15	
58	58	0	-	-	
38	43	-5	6	-	-6
61	53	8	8	8	
56	52	4	4.5	4.5	
56	57	-1	1	-	-1
34	44	-10	9	-	-9
55	57	-2	2.5	-	-2.5
65	40	25	14	14	
75	68	7	7	7	
Total				101.5	-18.5

Calculated value of T = 18.5

Table value of Wilcoxon's T table = 25

As the calculated value is less than the table value we accept the null hypothesis. i.e., there is no significant difference between the preference of machines A and B.

Case :2

When the number of matched pairs are more than 25

Procedure:-

1. Set up null hypothesis:

   H_0: There is no significant difference.

   H_1: There is significant difference.
2. Find the difference between each pair of values.
3. Assign ranks to the differences from the smallest to the largest without any regard to sign.
4. Then actual signs of each difference are put to the corresponding ranks.

5. Find the total of positive ranks and negative ranks.
6. Apply Z test and compute the value of 'Z'

$$Z = \frac{T-U}{\sigma}$$

Where T = Smaller value as per steps (5)

$$U = \frac{n(n+1)}{4}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

7. Obtain table value of Z at specified level of significance for infinity degrees of freedom.
8. Decide whether to accept or reject the null hypothesis.

CHAPTER 11

ANALYSIS OF VARIANCE

Definition of Analysis of Variance

Analysis of variance may be defined as a technique which analyses the variance of two or more comparable series (or samples) for determining the significance of differences in their arithmetic means and for determining whether different samples under study are drawn from same population or not, with the of the statistical technique, called F – test.

Characteristics of Analysis of Variance:

1. It makes statistical analysis of variance of two or more samples.
2. It tests whether the difference in the means of different sample is due to chance or due to any significance cause.
3. It uses the statistical test called, F – Ratio.

Types of Variance Analysis:

There are two types of variance analysis. They are:-

1. One way Analysis of Variance
2. Two way analysis of Variance

One way Analysis of Variance:

In one way analysis of variance, observations are classified into groups on the basis of a single criterion. For example, yield of a crop is influenced by quality of soil, availability of rainfall, quantity of seed, use of fertilizer, etc. It we study the influence of one factor, It is called one way analysis of variance.

If we want to study the effect of fertilizer of yield of crop, we apply different kinds of fertilizers on different paddy fields and try to find out the difference in the effect of these different kinds of fertilizers on yield.

Procedure:-

1.Set up null and alternative hypothesis:

H_0: There is no significant difference.

H_1: There is significant difference.

2. Compute sum of squares Total (SST)

$$\text{SST} = \text{Sum of squares of all observations} - \frac{T^2}{N}$$

3. Compute sum of squares between samples (SSC)

$$\text{SSC} = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} + \ldots\ldots\ldots - \frac{T^2}{N}$$

4. Compute sum of squares within sample (SSE)

$$\text{SSE} = \text{SST} - \text{SSC}$$

5. Compute MSC

$$\text{MSC} = \frac{\text{SSC}}{\text{d.f.}} = \frac{\text{SSC}}{\text{C}-1}$$

6. Compute MSE

$$\text{MSE} = \frac{\text{SSE}}{\text{d.f.}} = \frac{\text{SSE}}{\text{C}-1}$$

7. Compute F – ratio:

$$\text{F} = \frac{\text{Larger Variance}}{\text{Smaller Variance}}$$

8. Incorporate all these in an ANOVA TABLE as flows:

ANOVA TABLE				
Source of Variation	Sum of Squares	Degree of freedom	Means square	F - Ratio
Between Samples	SSC	C-1	$\text{MSC} = \frac{\text{SSC}}{\text{C}-1}$	$\text{F} = \frac{\text{Larger Variance}}{\text{Smaller Variance}}$
Within Sample	SSE	N-C	$\text{MSE} = \frac{\text{SSE}}{\text{C}-1}$	
Total	SST	N-1		

9. Obtain table value at corresponding to the level of significance and for degree of freedom of (C-1, N-C).

10. Decide whether to accept or reject the null hypothesis.

Qn: Given below are the yield (in Kg.) per acre for 5 trial plots of 4 varieties of treatments.

	Treatment			
Plot name	**1**	**2**	**3**	**4**
A	42	48	68	80
B	50	66	52	94
C	62	68	76	78
D	34	78	64	82
E	52	70	70	66

Carry out an analysis of variance and state whether there is any significant difference in treatments.

Sol: H_0: There is no significant difference in treatments.

H_1: There is significant difference in treatments.

X_1	X_2	X_3	X_4	X_1^2	X_2^2	X_3^2	X_4^2
42	48	68	80	1764	2304	4624	6400
50	66	52	94	2500	4356	2704	8836
62	68	76	78	3844	4624	5776	6084
34	78	64	82	1156	6084	4096	6724
52	70	70	66	2704	4900	4900	4356
$\Sigma X_1 = 240$	$\Sigma X_2 = 330$	$\Sigma X_3 = 330$	$\Sigma X_4 = 400$	11,968	22,268	22,100	32,400

SST = Sum of squares of all items - $\frac{T^2}{N}$

$= (11,968+22,268+22,100+32,400) - \frac{(240+330+330+400)^2}{20}$

$= 88,736 - \frac{1300^2}{20}$

$= 88,736 - \frac{16,90,000}{20}$

$= 88,736 - 84,500 = 4,236$

$SSC = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N} + \frac{(\sum X_4)^2}{N} - \frac{T^2}{N}$

$= \frac{240^2}{5} + \frac{330^2}{5} + \frac{330^2}{5} + \frac{400^2}{5} - \frac{1300^2}{20}$

$= 11,520+21,780+21,780+32,000 - 84,500$

$= 87,080 - 84,500 = 2,580$

ONE WAY ANOVA TABLE				
Source of Variation	Sum of Squares	Degree of freedom	Means square	F - Ratio
Between Samples	2,580	C-1= 3	MSC = $\frac{2580}{3}$ =860	
Within Sample	1,656	N-C = 16	MSE = $\frac{1656}{16}$ =103.5	F = $\frac{860}{103.5}$ = 8.31
Total	4,236	N-1= 19		

Calculated value of F is 8.31.

Table value of F at 5% level of significance for (3.16) degree of freedom is 3.24.

As the calculated value is greater than the table value, we reject the null hypothesis. We can conclude that there is significant difference in treatments. In other words, treatments do not have the same effect.

Qn: The following data relate to the yield of 4 varieties of rice each shown on 5 plots. Find whether there is significant difference between the mean yield of these varieties.

	Treatment			
Plot name	**1**	**2**	**3**	**4**
P	99	103	109	104
Q	101	102	103	100
R	103	100	107	103
S	99	105	97	107
T	98	95	99	106

Sol: Apply coding method. Subtract 100 from all the observations.

X_1 (A)	X_2 (B)	X_3 (C)	X_4 (D)	X_1^2	X_2^2	X_3^2	X_4^2
-1	3	9	4	1	9	81	16
1	2	3	0	1	4	9	0
3	0	7	3	9	0	49	9
-1	5	-3	7	1	25	9	49
-2	-5	-1	6	4	25	1	36
$\Sigma X_1 = 0$	$\Sigma X_2 = 5$	$\Sigma X_3 = 15$	$\Sigma X_4 = 20$	16	63	149	110

H_0: There is no significant difference between the mean yield of different varieties.

H_1: There is significant difference between mean yield of varieties.

SST = Sum of squares of all items - $\frac{T^2}{N}$

$= (16+63+149+110) - \frac{(0+5+15+20)^2}{20}$

$= 338 - \frac{40^2}{20}$

$= 338 - \frac{1600}{20}$

$= 338 - 80 = \underline{\underline{258}}$

$SSC = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} + \frac{(\sum X_4)^2}{N_4} - \frac{T^2}{N}$

$= \frac{0^2}{5} + \frac{5^2}{5} + \frac{15^2}{5} + \frac{20^2}{5} - \frac{40^2}{20}$

$= \frac{0}{5} + \frac{25}{5} + \frac{225}{5} + \frac{400}{5} - \frac{1600}{20}$

$= 0+5+45+80 - 80$

$= \underline{\underline{50}}$

ONE WAY ANOVA TABLE				
Source of Variation	Sum of Squares	Degree of freedom	Means square	F - Ratio
Between Samples	SSC = 50	C-1= 3	MSC = $\frac{50}{3}$ = 16.67	
Within Sample	SSE = 208	N-C = 16	MSE = $\frac{208}{16}$ =13	F = $\frac{16.67}{13}$ = 1.28
Total	SST = 258	N-1= 19		

Calculated value of F is 1.28

Degree of freedom is (3.16)

Table value at 5% level of significance and (3.16) d.f. is 3.24.

As the calculated value is less than the table value, we accept the null hypothesis.

∴ There is no significant difference between the mean yield of these varieties.

TWO WAY ANALYSIS OF VARIANCE

Two way analysis of variance is used to test the effect of two factors simultaneously on a particular variable.

Procedure:-

1. Set up null and alternative hypothesis.

   H_0: There is no significant difference between columns.
   There is no significant difference between rows.

   H_1: There is significant difference between columns.
   There is significant difference between rows.

2. Compute SST

   SST = Sum of squares of all observations - $\frac{T^2}{N}$

3. Compute SSC

$$SSC = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} + \dots\dots\dots - \frac{T^2}{N}$$

4. Compute SSR

$$SSR = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} + \dots\dots\dots - \frac{T^2}{N}$$

Here $\sum X_1, \sum X_2$, etc denote the row totals

5. Compute SSE

SSE = SST – (SSC + SSR)

6. Compute MSC

$$MSC = \frac{SSC}{d.f.} = \frac{SSC}{C-1}$$

7. Compute MSR

$$MSR = \frac{SSR}{d.f.} = \frac{SSR}{r-1}$$

8. Compute MSE

$$MSE = \frac{SSE}{d.f.} = \frac{SSE}{(C-1)(r-1)}$$

9. Compute F – ratio in respect of columns

$$F_c = \frac{MSC}{MSE}$$

10. Compute F – ratio in respect of rows

$$F_r = \frac{MSR}{MSE}$$

11. Obtain the table value

12. Decide whether to accept or reject the H_0:

TWO WAY ANOVA TABLE

Source of Variation	Sum of Squares	Degree of freedom	Means square	F - Ratio
Between Columns	SSC	c-1	$MSC = \frac{SSC}{c-1}$	$F_c = \frac{MSC}{MSE}$
Between Rows	SSR	r-1	$MSE = \frac{SSR}{r-1}$	
Residual	SSE	(c-1)(r-1)	$MSE = \frac{SSE}{(c-1)(r-1)}$	$F_r = \frac{MSR}{MSE}$
Total	SST	N-1		

Qn: Apply the technique of analysis of variance to the following date relating to yields of 4 varieties of wheat in 3 blocks:

	Blocks		
Varieties	**X**	**Y**	**Z**
A	10	9	8
B	7	7	6
C	8	5	4
D	5	4	4

Carry two-way analysis of variance.

Sol:

Varieties	X X_1	Y X_2	Z X_3	Total	X_1^2	X_2^2	X_3^2	Total
A(X_1)	10	9	8	27	100	81	64	245
B (X_2)	7	7	6	20	49	49	36	134
C(X_3)	8	5	4	17	64	25	16	105
D(X_4)	5	4	4	13	25	16	16	57
Total	30	25	22	77	238	171	132	541

H_0: There is no significant difference between blocks.
There is no significant difference between varieties.

H_1: There is significant difference between block.
There is significant difference between varieties.

SST = Sum of squares of all items - $\frac{T^2}{N}$

$= 541 - \frac{77^2}{12}$

$= 541 - \frac{5929}{12}$

$= 541 - 494.083 = 46.917$

$SSC = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} - \frac{T^2}{N}$

$= \frac{30^2}{4} + \frac{25^2}{4} + \frac{22^2}{4} - \frac{77^2}{12}$

$$= \frac{900}{4} + \frac{625}{4} + \frac{484}{4} - \frac{5929}{12}$$

$$= 225+156.25+121 - 494.083$$

$$= 502.25 - 494.083 = \underline{\underline{8.167}}$$

$$SSR = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} + \frac{(\sum X_4)^2}{N_4} - \frac{T^2}{N}$$

$$= \frac{27^2}{3} + \frac{20^2}{3} + \frac{17^2}{3} + \frac{13^2}{3} - \frac{77^2}{12}$$

$$= \frac{729}{3} + \frac{400}{3} + \frac{289}{3} + \frac{169}{3} - \frac{5929}{12}$$

$$= 243+133.333+96.333+56.333 - 494.083$$

$$= \underline{\underline{34.916}}$$

TWO WAY ANOVA TABLE				
Source of Variation	Sum of Squares	Degree of freedom	Means square	F - Ratio
Between Columns	SSC= 8.167	c-1= 2	MSC = $\frac{8.167}{2}$ =4.084	$F_c = \frac{4.084}{0.639}$
Between Rows	SSR=34.91	r-1= 3	MSE = $\frac{34.916}{3}$ = 11.639	= 6.39
Residual	SSE= 3.834	(c-1)(r-1)=6	MSE = $\frac{3.834}{6}$ =0.639	$F_r = \frac{11.639}{0.639}$ = 18.21
Total	SST= 46.917	N-1= 11		

Between columns (blocks):-

Degree of freedom = (2, 6)

Calculated Value = 6.39

Table Value = 5.1433

As the calculated value is more than the table value, we reject the null hypothesis. It is concluded that there is significant difference between blocks.

i.e., the mean productivity between blocks are not same.

Between rows (varieties):-

Degree of freedom = (3.6)

Calculated value = 18.21

Table value = 4.7571

As the calculated value is greater than the table value, we reject the null hypothesis. This means that there is significant difference in mean productivity of the varieties.

Qn: The following date presents the number of units of production per day turned out by 5 different workers using 4 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

(a) Test whether the mean productivity is the same for the different machine types.

(b) Test whether the 5 workers differ with respect to mean productivity.

Let us apply coding method. Let us subtract 40 from all the observations.

Workers	A X_1	B X_2	C X_3	D X_4	Total	X_1^2	X_2^2	X_3^2	X_4^2	Total
1(X_1)	4	-2	7	-4	5	16	4	49	16	85
2 (X_2)	6	0	12	3	21	36	0	144	9	189
3(X_3)	-6	-4	4	-8	-14	36	16	16	64	132
4(X_4)	3	-2	6	-7	0	9	4	36	49	98
5(X_5)	-2	2	9	-1	8	4	4	81	1	90
Total	5	-6	38	-17	20	101	28	326	139	594

H_0: There is no significant difference in the mean productivity of machine type.

There is no significant difference in the mean productivity of workers.

H_1: There is significant difference between in the mean productivity of machine type.

There is significant difference between in the mean productivity of workers.

SST = Sum of squares of all items - $\frac{T^2}{N}$

$= 594 - \frac{20^2}{20}$

$= 594 - \frac{400}{20} = 594 - 20 = 574$

$SSC = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} + \frac{(\sum X_4)^2}{N_4} - \frac{T^2}{N}$

$= \frac{5^2}{5} + \frac{(-6)^2}{5} + \frac{38^2}{5} + \frac{(17)^2}{5} - \frac{20^2}{20}$

$= \frac{25}{5} + \frac{36}{5} + \frac{1444}{5} + \frac{289}{5} - \frac{400}{20}$

$= \frac{25+36+76+289}{5} - \frac{400}{20}$

$= \frac{1794}{5} - 20 = 358.8 - 20 = 338.8$

$SSR = \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \frac{(\sum X_3)^2}{N_3} + \frac{(\sum X_4)^2}{N_4} + \frac{(\sum X_5)^2}{N_5} - \frac{T^2}{N}$

$= \frac{5^2}{4} + \frac{21^2}{4} + \frac{-14^2}{4} + \frac{0^2}{4} + \frac{8^2}{4} - \frac{20^2}{20}$

$= \frac{25+441+196+64}{4} - 20$

$= \frac{726}{4} - 20$

$= 181.5 - 20$

$= 161.50$

TWO WAY ANOVA TABLE

Source of Variation	Sum of Squares	Degree of freedom	Means square	F - Ratio
Between Samples	SSC= 338.8	c-1= 3	MSC = $\frac{338.8}{3}$ =112.93	$F_c = \frac{112.93}{6.142}$ = 18.39
Between Rows	SSR=161.5	r-1= 4	MSE = $\frac{161.5}{4}$ = 40.375	$F_r = \frac{40.375}{6.142}$ = 6.57
Residual	SSE= 73.7	(c-1)(r-1)=12	MSE = $\frac{73.7}{12}$ = 6.142	
Total	SST= 574.0	N-1= 19		

Between Columns (Machine type)

Calculated value = 18.39

Degree of freedom = (3.12)

Table value of F = 3.49

As the calculated value is greater than the table value, we reject the H_0. ∴ Mean productivity is not the same for different types of machines.

Between rows (workers):

Calculated Value = 6.57

Degree of freedom = (4.12)

Table value of F = 3. 2592

As the calculated value is greater than the table value, we reject the H_0

∴ Mean productivity is not the same for different workers.