Ramnath Guljarilal college of commerce

DEPARTMENT OF BUSINESS MANAGEMENT

LECTURE NOTES

ON

STATISTICS FOR MANAGEMENT

MBA – I SEMESTER

UNIT – I: INTRODUCTION TO STATISTICS

DEFINITION OF STATISTICS

Branch of mathematics concerned with collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability). Statistics can interpret aggregates of data too large to be intelligible by ordinary observation because such data (unlike individual quantities) tend to behave in regular, predictable manner. It is subdivided into descriptive statistics and inferential statistics.

HISTORY OF STATISTICS

The Word statistics have been derived from Latin word ―Status‖ or the Italian word ―Statista‖, meaning of these words is ―Political State‖ or a Government. Shakespeare used a word Statist is his drama Hamlet (1602). In the past, the statistics was used by rulers. The application of statistics was very limited but rulers and kings needed information about lands, agriculture, commerce, population of their states to assess their military potential, their wealth, taxation and other aspects of government.

Gottfried Achenwall used the word statistik at a German University in 1749 which means that political science of different countries. In 1771 W. Hooper (Englishman) used the word statistics in his translation of Elements of Universal Erudition written by Baron B.F Bieford, in his book statistics has been defined as the science that teaches us what is the political arrangement of all the modern states of the known world. There is a big gap between the old statistics and the modern statistics, but old statistics also used as a part of the present statistics.

During the 18th century the English writer have used the word statistics in their works, so statistics has eveloped gradually during last few centuries. A lot of work has been done in the end of the nineteenth century.

At the beginning of the 20th century, William S Gosset was developed the methods for decision making based on small set of data. During the 20th century several statistician are active in developing new methods, theories and application of statistics. Now these days the availability of electronics computers is certainly a major factor in the modern development of statistics.

Descriptive Statistics and Inferential Statistics

Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

MANAGERIAL APPLICATIONS OF STATISTICS

Statistics is a the mathematical science involving the collection, analysis and interpretation of data. A number of specialties have evolved to apply statistical theory and methods to various disciplines. Certain topics have "statistical" in their name but relate to manipulations of probability distributions rather than to statistical analysis.

- **Actuarial science** is the discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries.
- **Astrostatistics** is the discipline that applies statistical analysis to the understanding of astronomical data.
- **Biostatistics** is a branch of biology that studies biological phenomena and observations by means of statistical analysis, and includes medical statistics.
- **Business analytics** is a rapidly developing business process that applies statistical methods to data sets (often very large) to develop new insights and understanding of business performance & opportunities

- **Chemometrics** is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.
- **Demography** is the statistical study of all populations. It can be a very general science that can be applied to any kind of dynamic population, that is, one that changes over time or space.
- **Econometrics** is a branch of economics that applies statistical methods to the empirical study of economic theories andrelationships.
- **Environmental statistics** is the application of statistical methods to environmental science. Weather, climate, air and water quality are included, as are studies of plant and animal populations.
- **Geostatistics** is a branch of geography that deals with the analysis of data from disciplines such as petroleum geology, hydrogeology, hydrology, meteorology, oceanography, geochemistry, geography.
- **Operations research** (or Operational Research) is an interdisciplinary branch of applied mathematics and formal science that uses methods such as mathematical modeling, statistics, and algorithms to arrive at optimal or near optimal solutions to complex problems.
- **Population ecology** is a sub-field of ecology that deals with the dynamics of species populations and how these populations interact with the environment.
- **Quality control** reviews the factors involved in manufacturing and production; it can make use of statistical sampling of product items to aid decisions in process control or in accepting deliveries.
- **Quantitative psychology** is the science of statistically explaining and changing mental processes and behaviors in humans.
- **Statistical finance**, an area of econophysics, is an empirical attempt to shift finance from its normative roots to a positivist framework using exemplars from statistical physics with an emphasis on emergent or collective properties of financial markets.
- **Statistical mechanics** is the application of probability theory, which includes mathematical tools for dealing with large populations, to the field of mechanics, which is concerned with the motion of particles or objects when subjected to a force.

Statistical physics is one of the fundamental theories of physics, and uses methods of probability theory in solving physical problems.

STATISTICS AND COMPUTERS

Crunch numbers to the nth degree — and see what happens. When you study computer science and mathematics, you_ll use algorithms and computational theory to create mathematical models or define formulas that solve mathematical problems. In other words, you'll design new tools that can predict the future.

The Computer Applications option gives students the flexibility to combine a traditional computer science degree with a non-traditional field. Our state-of-the- art labs for high-performance computing, networks and artificial intelligence will give you experience with the tools you'll use in the field. Through labs, lectures and projects, you'll also:

1. Investigate the computational limits of the algorithms and data structures that support complex software systems
2. Develop new applications and tools in multi-disciplinary areas of science and research
3. Explore opportunities for advanced computer modeling and simulation

IMPORTANCE OF STATISTICS IN DIFFERENT FIELDS

Statistics plays a vital role in every fields of human activity. Statistics has important role in determining the existing position of per capita income, unemployment, population growth rate, housing, schooling medical facilities etc…in a country. Now statistics holds a central position in almost every field like Industry, Commerce, Trade, Physics, Chemistry, Economics, Mathematics, Biology, Botany, Psychology, Astronomy etc…, so application of statistics is very wide. Now we discuss some important fields in which statistics is commonly applied.

1. **Business:** Statistics play an important role in business. A successful businessman must be very quick and accurate in decision making. He knows that what his customers wants, he should therefore, know what to produce and sell and in what quantities. Statistics helps businessman to plan production according to the taste of the costumers, the quality of the products can also be checked more efficiently by using statistical methods. So all the activities of the businessman based on statistical information. He can make correct decision about the location of business, marketing of the products, financial resources etc…

2. **In Economics:** Statistics play an important role in economics. Economics largely depends upon statistics. National income accounts are multipurpose indicators for the economists and administrators. Statistical methods are used for preparation of these accounts. In economics research statistical methods are used for collecting and analysis the data and testing hypothesis. The relationship between supply and demands is studies by statistical methods, the imports and exports, the inflation rate, the per capita income are the problems which require good knowledge of statistics.

3. **In Mathematics:** Statistical plays a central role in almost all natural and social sciences. The methods of natural sciences are most reliable but conclusions draw from them are only probable, because they are based on incomplete evidence. Statistical helps in describing these measurements more precisely. Statistics is branch of applied mathematics. The large number of statistical methods like probability averages, dispersions, estimation etc… is used in mathematics and different techniques of pure mathematics like integration, differentiation and algebra are used in statistics.

4. **In Banking:** Statistics play an important role in banking. The banks make use of statistics for a number of purposes. The banks work on the principle that all the people who deposit their money with the banks do not withdraw it at the same time. The bank earns profits out of these deposits by lending to others on interest. The bankers use statistical approaches based on probability to estimate the numbers of depositors and their claims for a certain day.

5. **In State Management (Administration):** Statistics is essential for a country. Different policies of the government are based on statistics. Statistical data are now widely used in taking all administrative decisions. Suppose if the government wants to revise the pay scales of employees in view of an increase in the living cost, statistical methods will be used to determine the rise in the cost of living. Preparation of federal and provincial government budgets mainly depends upon statistics because it helps in estimating the expected expenditures and revenue from different sources. So statistics are the eyes of administration of the state.

6. **In Accounting and Auditing:** Accounting is impossible without exactness. But for decision making purpose, so much precision is not essential the decision may be taken on the basis of approximation, know as statistics. The correction of the values of current asserts is made on the basis of the purchasing power of money or the current value of it. In auditing sampling techniques are commonly used. An auditor determines the sample size of the book to be audited on the basis of error.

7. **In Natural and Social Sciences:** Statistics plays a vital role in almost all the natural and social sciences. Statistical methods are commonly used for analyzing the experiments results, testing their significance in Biology, Physics, Chemistry, Mathematics, Meteorology, Research chambers of commerce, Sociology, Business, Public Administration, Communication and Information Technology etc…

8. **In Astronomy:** Astronomy is one of the oldest branches of statistical study; it deals with the measurement of distance, sizes, masses and densities of heavenly bodies by means of observations. During these measurements errors are unavoidable so most probable measurements are founded by using statistical methods.

MEASURES OF CENTRAL TENDENCY

MEASURES OF CENTRAL TENDENCY:

The term **central tendency** refers to the "middle" value or perhaps a typical value of the data, and is measured using the **mean**, **median**, or **mode**. Each of these measures is calculated differently, and the one that is best to use depends upon the situation.

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations. There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

Arithmetic mean or mean:

Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. It is denoted by the symbol If the variable x assumes n values x1, x2 … xn then the mean is given by

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" is used for the mean of a population. The symbol "M" is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \Sigma X/N$$

Where, ΣX is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for M is essentially identical:

M = ΣX/N

Where, ΣX is the sum of all the numbers in the sample and N is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Example

Calculate the mean for pH levels of soil 6.8, 6.6, 5.2, 5.6, 5.8

$$\bar{x} = \frac{6.8+6.6+5.2+5.6+5.8}{5} = \frac{30}{5} = 6$$

Grouped Data

The mean for grouped data is obtained from the following formula:

$$\bar{x} = \frac{\sum fx}{n}$$

Where x = the mid-point of individual class

f = the frequency of individual class

n = the sum of the frequencies or total frequencies in a sample.

Short-cut method

$$\bar{x} = A + \frac{\sum fd}{n} x\ c$$

Where $d = \frac{x - A}{c}$

A = any value in x

n = total frequency

c = width of the class interval

Example

Given the following frequency distribution, calculate the arithmetic mean

Marks : 64	63	62	61	60	59
Number of Students : 8	18	12	9	7	6

Solution

X	**F**	**Fx**	**D=x-A**	**Fd**
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	-1	-9
60	7	420	-2	-14
59	6	354	-3	-18
	60	3713		-7

Short-cut method

$$\bar{x} = A + \frac{\sum fd}{n} x\ c$$

Here A = 62

$$\bar{x} = 62 - \frac{7}{60} x\ 1 = 61.88$$

Example

For the frequency distribution of seed yield of plot given in table, calculate the mean yield per plot.

Yield per plot in(ing)	**64.5-84.5**	**84.5-104.5**	**104.5-124.5**	**124.5-144.5**
No of plots	3	5	7	20

Solution

Yield (in g)	**No of Plots (f)**	**Mid X**	$d = x - A$	**Fd**
64.5-84.5	3	74.5	-1	-3
84.5-104.5	5	94.5	0	0
104.5-124.5	7	114.5	1	7
124.5-144.5	20	134.5	2	40
Total	**35**			**44**

A=94.5

The mean yield per plot is Direct method:

$$\bar{x} = \frac{\sum fx}{n} = \frac{(74.5 \times 3) + (94.5 \times 5) + (114.5 \times 7) + (134.5 \times 20)}{35} = \frac{4187.5}{35} = 119.64 \text{ gms}$$

Shortcut method

$$\bar{x} = A + \frac{\sum fd}{n} x\ c$$

$$\bar{x} = 94.5 + \frac{44}{35} x\ 20 = 119.64\ g$$

Merits and demerits of Arithmetic mean

Merits

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

Median

The median is the middle most item that divides the group into two equal parts, one part comprising all values greater, and the other, all values less than that item.

Ungrouped or Raw data

Arrange the given values in the ascending order. If the numbers of values are odd, median is the middle value. If the numbers of values are even, median is the mean of middle two values. By formula $\left(\frac{n+1}{2}\right)^{th}$

When n is odd, Median = Md $value$

When n is even, Average of $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)^{th}$ $value$

Example

If the weights of sorghum ear heads are 45, 60,48,100,65 gms, calculate the median

Solution

Here n = 5

First arrange it in ascending order 45, 48, 60, 65, 100

Median = $\left(\frac{n+1}{2}\right)^{th} value$

= $\left(\frac{5+1}{2}\right) =$ $value\ 3^{rd}$ =60

Example

If the sorghum ear- heads are 5,48, 60, 65, 65, 100 gms, calculate the median.

Solution

Here n = 6

$$\text{Median} = \text{Average of} \left(\frac{n}{2}\right) and \left(\frac{n}{2}+1\right)^{th} value$$

$$\left(\frac{n}{2}\right) = \frac{6}{2} = 3^{rd} value = 60 \quad \text{and} \left(\frac{n}{2}+1\right) = \frac{6}{2}+1 = 4^{th} value = 65$$

$$\text{Median} = \frac{60+65}{2} = 62.5\,g$$

Grouped data

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series

Step1: Find cumulative frequencies.

Step2 : Find $\left(\frac{n}{2}+1\right)$ $\left(\frac{n}{2}+1\right)$

Step3: See in the cumulative frequencies the value just greater than Step4: Then the

Example

The following data pertaining to the number of insects per plant. Find median number of insects per plant.

Number of insects per plant (x)	1	2	3	4	5	6	7	8	9	10	11	12
No. of plants(f)	1	3	5	6	10	13	9	5	3	2	2	1

Solution

Form the cumulative frequency table

x	**f**	**cf**
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
	60	

Median = size of $\left(\frac{n+1}{2}\right)^{th}$ item

Here the number of observations is even. Therefore median = average of (n/2)th item and (n/2+1)th item.

= (30^{th} item +31^{st} item) / 2 = (6+6)/2 = 6

Hence the median size is 6 insects per plant.

Continuous Series

The steps given below are followed for the calculation of median in continuous series. Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{n}{2}\right)$

Step3: See in the cumulative frequency the value first greater than $\left(\frac{n}{2}\right)$, Then the corresponding

class interval is called the Median class. Then apply the formula

Median $= l + \frac{\frac{n}{2} - m}{f} \times c$ _____

where l = Lower limit of the medianal class

m = cumulative frequency preceding the medianal class c = width of the class

f =frequency in the median class.

n=Total frequency.

Example

For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the median.

Weights of ear heads (in g)	No of ear heads (f)	Less than class	Cumulative frequency (m)
60-80	22	<80	22
80-100	38	<100	60
100-120	45	<120	105
120-140	35	<140	140
140-160	24	<160	164
Total	164		

Solution

$$\text{Median} = l + \frac{\frac{n}{2} - m}{f} \times c$$

$$\left(\frac{n}{2}\right) \quad \left(\frac{164}{2}\right) = 82$$

It lies between 60 and 105. Corresponding to 60 the less than class is 100 and corresponding to 105 the less than class is 120. Therefore the medianal class is 100-120. Its lower limit is 100. Here 100, n=164 , f = 45 , c = 20, m =60

$$\text{Median} = 100 + \frac{82 - 60}{45} \times 20 = 109.78 \ gms$$

Merits of Median

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.

Demerits of Median

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in calculating mean deviation.
4. It does not take into account all the observations.

Mode

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it. It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in agriculture like to find typical height of a crop variety, maximum source of irrigation in a region, maximum disease prone paddy variety. Thus the mode is an important measure in case of qualitative data.

Computation of the mode Ungrouped or Raw Data

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example

Find the mode for the following seed weight 2 , 7, 10, 15, 10, 17, 8, 10, 2 gms

Mode = 10

In some cases the mode may be absent while in some cases there may be more than one mode.

Example 9

(1) 12, 10, 15, 24, 30 (no mode)

(2) 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

the modal values are 7 and 10 as both occur 3 times each.

Grouped Data

For Discrete distribution, see the highest frequency and corresponding value of x is mode.

Example:

Find the mode for the following

Weight of sorghum in gms (x)	No. of ear head(f)
50	4
65	6
75	16
80	8
95	7
100	4

Solution

The maximum frequency is 16. The corresponding x value is 75.

□ mode = 75 gms.

Continuous distribution

Locate the highest frequency the class corresponding to that frequency is called the modal class. Then apply the formula.

Mode = $l + \frac{f_s}{f_p + f_s} \times c$

Where l lower limit of the model class

f_p = the frequency of the class preceding the model class f_s = the frequency of the class succeeding the modelclass

and c = class interval

Example

For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the mode

Weights of ear heads (g)	No of ear heads (f)	
60-80	22	
80-100	38	f_p
100-120	45	f
120-140		f_s
140-160	20	
Total	**160**	

Solution

Mode = $l + \frac{f_s}{f_p + f_s} \times c$

Here l = 100, f = 45, c = 20, m =60, f_p =38, f_s =35

Mode = $100 + \frac{35}{38 + 35} \times 20$

$= 100 + \frac{35}{73} \times 20$

= 109.589

Geometric mean

The geometric mean of a series containing n observations is the nth root of the product of the values. If x1, x2…, xn are observations then

G.M= $\sqrt[n]{x_1, x_2 ... x_n}$

= $(x_1, x_2 ... x_n)^{1/n}$

Log GM = $\frac{1}{n}\log(x_1, x_2 ... x_n)$

= $\frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n)$

= $\frac{\sum \log x_i}{n}$

GM = Antilog For grouped data $\frac{\sum \log x_i}{n}$

GM = Antilog $\left[\frac{\sum f_i \log x_i}{n}\right]$

GM is used in studies like bacterial growth, cell division, etc.

Example

If the weights of sorghum ear heads are 45, 60, 48,100, 65 gms. Find the Geometric mean for the following data

Weight of ear head x (g)	**Log x**
45	1.653
60	1.778
48	1.681
100	2.000
65	1.813
Total	**8.925**

Solution

Here n = 5 GM = Antilog $\frac{\sum \log x_i}{n}$

= Antilog $\frac{8.925}{5}$

= Antilog 1.785

Grouped Data

Example

Find the Geometric mean for the following

Weight of sorghum (x)	No. of ear head(f)
50	4
65	6
75	16
80	8
95	7
100	4

Solution

Weight of sorghum (x)	No. of ear head(f)	Log x	f x log x
50	5	1.699	8.495
63	10	10.799	17.99
65	5	1.813	9.065
130	15	2.114	31.71
135	15	2.130	31.95
Total	**50**	**9.555**	**99.21**

Here n= 50 GM = Antilog

$$= \text{Antilog} \left[\frac{\sum f \log x_i}{\left[\frac{99.21}{50} \right]_n} \right]$$

= Antilog 1.9842 = 96.43

Continuous distribution Example

For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the Geometric mean

Weights of ear heads (in g)	No of ear heads (f)
60-80	22
80-100	38
100-120	45
120-140	35
140-160	20
Total	**160**

Solution

Weights of ear heads (in g)	No of ear heads (f)	Mid x	Log x	f log x
60-80	22	70	1.845	40 59
80-100	38	90	1.954	74.25
100-120	45	110	2.041	91.85
120-140	35	130	2.114	73.99
140-160	20	150	2.176	43.52
Total	**160**			**324.2**

Here n = 160 GM = Antilog

$$= \text{Antilog} \left[\frac{\sum f \log x_i}{\frac{324.2}{160}} \right]$$

$$= \text{Antilog} \left[2.02625 \right]$$

= 106.23

Harmonic mean (H.M)

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x1, x2…..xn are n observations,

$$\text{H.M} = \frac{n}{\sum_{i=n}^{n}\left(\frac{1}{x_i}\right)}$$

For a frequency distribution

$$\text{H.M} = \frac{n}{\sum_{i=n}^{n} f\left(\frac{1}{x_i}\right)}$$

H.M is used when we are dealing with speed, rates, etc.

Example

From the given data 5, 10,17,24,30 calculate H.M.

X	$\frac{1}{x}$
5	0.2000
10	0.1000
17	0.0588
24	0.0417
30	0.4338

$$\text{H.M} = \frac{5}{0.4338} = 11.526$$

Example

Number of tomatoes per plant are given below. Calculate the harmonic mean.

Number of tomatoes per plant	20	21	22	23	24	25
Number of plants	4	2	7	1	3	1

Solution

Number of tomatoes per plant (x)	No of plants(f)	$\frac{1}{x}$	$f\left(\frac{1}{x}\right)$
20	4	0.0500	0.2000
21	2	0.0476	0.0952
22	7	0.0454	0.3178
23	1	0.0435	0.0435
24	3	0.0417	0.1251
25	1	0.0400	0.0400
	18		0.8216

$$\sum f\left(\frac{1}{x_i}\right) \qquad = \frac{0.1968}{} = 21.91$$

Merits of H.M

1. It is rigidly defined.
2. It is defined on all observations.
3. It is amenable to further algebraic treatment.
4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

Demerits of H.M

1. It is not easily understood.
2. It is difficult to compute.
3. It is only a summary figure and may not be the actual item in the series
4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.
5. It is rarely used in grouped data.

Percentiles

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The x^{th} percentile is that value below which x percent of values in the distribution fall. It may be noted that the median is the 50^{th} percentile.

For raw data, first arrange the n observations in increasing order. Then the x^{th} percentile is given by

$$P_x = \left\{\frac{x(n+1)}{100}\right\}^{th} \text{item}$$

For a frequency distribution the x^{th} percentile is given by

$$P_x = l + \left(\frac{(x.n/100) - cf}{f} \times C\right)$$

Where

= lower limit of the percentile calss which contains the x^{th} percentile value (x. n /100) cf = cumulative frequency uotp l

f = frequency of the percentile class

f C= class interval

N= total number of observations

Percentile for Raw Data or Ungrouped Data

Example

The following are the paddy yields (kg/plot) from 14 plots:
30,32,35,38,40.42,48,49,52,55,58,60,62,and 65 (after arranging in ascending order). The computation of 25^{th} percentile (Q_1) and 75^{th} percentile (Q_3) are given below:

$$P_{25}(or\ Q_1) = \left(\frac{25(14+1)}{100}\right)^{th} \text{ item}$$

$$= \left(3\frac{3}{4}\right)^{th} \text{ item}$$

$= 3^{rd}$ item + (4^{th} item – 3^{rd} item) $\left(\frac{3}{4}\right)$

= 35 + (38-35) $\left(\frac{3}{4}\right)$

= 35 + 3 $\left(\frac{3}{4}\right)$ = 37.25 kg

$$P_{75}(or\ Q_3) = \left(\frac{75(14+1)}{100}\right)^{th} \text{ item}$$

$$= \left(11\frac{1}{4}\right)^{th} \text{ item}$$

$= 11^{th}$ item + (12^{th} item – 11^{th} item) $\left(\frac{1}{4}\right)$

= 55 +(58-55) $\left(\frac{1}{4}\right)$

= 55 + 3 $\left(\frac{1}{4}\right)$ = 55.75 kg

Example

The frequency distribution of weights of 190 sorghum ear-heads are given below. Compute 25th percentile and 75th percentile.

Weight of ear-heads (in g)	No of ear heads
40-60	6
60-80	28
80-100	35
100-120	55
120-140	30
140-160	15
160-180	12
180-200	9
Total	**190**

Solution

Weight of ear-heads (in g)	No of ear heads	Less than class	Cumulative frequency	
40-60	6	< 60	6	
60-80	28	< 80	34	
47.5 80-100	35	<100	69	→
100-120	55	<120	124	
142.5 120-140	30	<140	154	→
140-160	15	<160	169	
160-180	12	<180	181	
180-200	9	<200	190	
Total	**190**			

or P_{25}, first find out $\left(\frac{25(190)}{100}\right)$, and for P_{75}, $\left(\frac{75(190)}{100}\right)$, and proceed as in the case of median.

For P_{25}, we have $\left(\frac{25(190)}{100}\right)$ = 47.5.

The value 47.5 lies between 34 and 69. Therefore, the percentile class is 80-100. Hence,

$$P_{25} = Q_1 = l + \left(\frac{(25.n/100) - cf}{f} \times C\right)$$

$$= 80 + \left(\frac{(47.5) - 34}{35} \times 20\right)$$

$$= 80 + \left(\frac{(13.5)}{35} \times 20\right)$$

= 80 +7.71 or 87.71 g.

Quartiles

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower).quartile (Q1) marks off the first one-fourth, the third (upper) quartile (Q3) marks off the three-fourth. It may be noted that the second quartile is the value of the median and 50th percentile.

Raw or ungrouped data

First arrange the given data in the increasing order and use the formula for Q1 and Q3 then quartile deviation, Q.D is given by

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Where $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ item and $Q_3 = 3\left(\frac{n+1}{4}\right)^{th}$ item

Example 18

Compute quartiles for the data given below (grains/panicles) 25, 18, 30, 8, 15, 5, 10, 35, 40, 45

Solution

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

$$Q_1 = \left(\frac{n+1}{4}\right)^{th}$$

$$= \left(\frac{10+1}{4}\right)^{th}$$

= (2.75)th item

= 2nd item + $\left(\frac{3}{4}\right)$ (3rd item – 2nd item)

= 8+ $\frac{3}{4}$ (10-8)

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

= 8+1.5

= 9.5

$Q_3 . = 3\left(\frac{n+1}{4}\right)^{th}$

= 3 x (2.75) th item

= (8.75)th item

= 8th item + $\left(\frac{1}{4}\right)$ (9th item – 8th item)

= 35+ $\frac{1}{4}$ (40-35)

= 35+1.25

= 36.25

Discrete Series

Step1: Find cumulative frequencies. Step2: Find

$\left(\frac{n+1}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{n+1}{4}\right)$, then the

corresponding value of *x* is Q1 Step4: Find

$3\left(\frac{n+1}{4}\right)$

Step5: See in the cumulative frequencies, the value just greater than $3\left(\frac{n+1}{4}\right)$,then the

corresponding value of x is Q3

Example

Compute quartiles for the data given bellow (insects/plant).

X	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

Solution

x	f	cf
5	4	4
8	3	7
12	2	9
15	4	13
19	3	16
24	4	20

$$\left(\frac{n+1}{4}\right)^{th} item = \left(\frac{24+1}{4}\right)^{th} item = \left(\frac{25}{4}\right) = 6.25^{th} item$$

$$Q_3 = 3\left(\frac{n+1}{4}\right)^{th} item = 3\left(\frac{24+1}{4}\right)$$ =18.75th item Q1=8; Q3=24

$Q_1 =$

Continuous series

Step1: Find cumulative frequencies Step2: Find

$\left(\frac{n}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{n}{4}\right)$, then the

corresponding class interval is called first quartile class.

Step4: Find $3\left(\frac{n}{4}\right)$ See in the cumulative frequencies the value just greater than $3\left(\frac{n}{4}\right)$ then the

corresponding class interval is called 3rd quartile class. Then apply the respective formulae

$$Q_1 = l_1 + \frac{\frac{n}{4} - m_1}{f_1} \times c_1$$

Q_3

Where l_1 = lower limit of the first quartile class

$$= l_3 + \frac{3\left(\frac{n}{4}\right) - m_3}{f_3} \times c_3$$

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = c.f. preceding the first quartile class

l_3 = 1ower limit of the 3rd quartile class f_3 = frequency of the 3rd quartile class c_3 = width of the 3rd quartile class

m_3 = c.f. preceding the 3rd quartile class

Table 1 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$\mu = \Sigma X/N$

$= 634/31$

$= 20.4516$

Table 1. Number of touchdown passes.

37 33 33 32 29 28 28 23 22 22 22 21 21 21 20 20 19 19 18 18 18 18 16 15 14 14 14 12 12 9 6

Although the arithmetic mean is not the only "mean" (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

Median

The median is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table 1, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile.

Computation of the Median

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is (4+7)/2 = 5.5. When there are numbers with the same values, then the formula for the third definition of the 50th percentile should be used.

Mode

The mode is the most frequently occurring value. For the data in Table 1, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables).

Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2 shows a grouped frequency distribution for the target res onse time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

GEOMETRIC MEAN

Geometric Mean is a special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc.

Example: What is the Geometric Mean of 2 and 18?

- First we multiply them: $2 \times 18 = 36$
- Then (as there are two numbers) take the square root: $\sqrt{36} = \mathbf{6}$

In one line:

Geometric Mean of 2 and 18 = $\sqrt{(2 \times 18)} = 6$

It is like the area is the same!

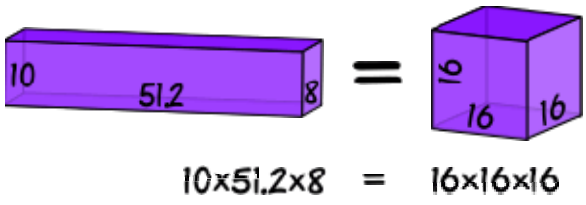
Example: What is the Geometric Mean of 10, 51.2 and 8?

- First we multiply them: $10 \times 51.2 \times 8 = 4096$
- Then (as there are three numbers) take the cube root: $\sqrt[3]{4096} = \mathbf{16}$

In one line:

Geometric Mean = $\sqrt[3]{(10 \times 51.2 \times 8)} = 16$

It is like the volume is the same:



Example: What is the Geometric Mean of 1,3,9,27 and 81?

- First we multiply them: $1 \times 3 \times 9 \times 27 \times 81 = 59049$
- Then (as there are 5 numbers) take the 5th root: $\sqrt[5]{59049} = \mathbf{9}$

In one line:

Geometric Mean = $\sqrt[5]{(1 \times 3 \times 9 \times 27 \times 81)} = 9$

I can't show you a nice picture of this, but it is still true that:

$1 \times 3 \times 9 \times 27 \times 81 = 9 \times 9 \times 9 \times 9 \times 9$

Harmonic Mean

A kind of average. To find the harmonic mean of a set of *n* numbers, add the reciprocals of the numbers in the set, divide the sum by *n*, then take the reciprocal of the result. The harmonic mean of $\{a1, a2, a3, a4, \ldots, an\}$ is given below.

Formula: $$\text{Harmonic Mean} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \frac{1}{a_4} + \cdots + \frac{1}{a_n}}$$

Example: For the numbers 4 and 9,

$$\text{Harmonic Mean} = \frac{2}{\frac{1}{4} + \frac{1}{9}} = \frac{72}{13} = 5.54$$

RANGE

The difference between the lowest and highest values.

In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9, so the range is 9 − 3 = 6. Range can also mean all the output values of a function.

Quartile Deviation :

In a distribution, partial variance between the upper quartile and lower quartile is known as 'quartile deviation'. Quartile Deviation is often regarded as semi inter quartile range.

Formula : (Upper quartile- lower quartile) / 2 upper quartile = 400, lower quartile = 200 then Quartile deviation (QD) = (400-200)/2 = 200/2

=100.

Mean Deviation

The mean of the distances of each value from their mean.

Yes, we use "**mean**" twice: Find the mean ... use it to work out distances ... then find the mean of those distances!

Three steps:

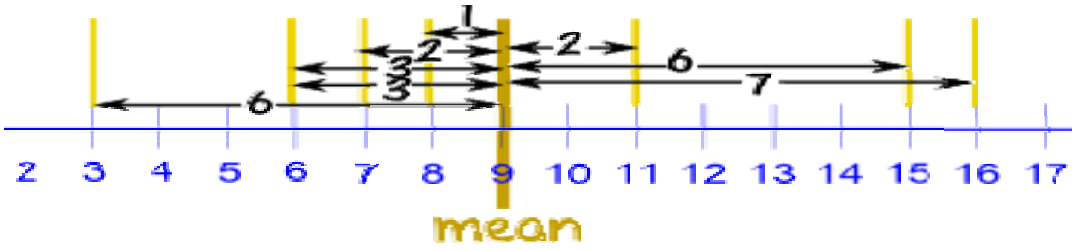
- 1. Find the mean of all values
- 2. Find the **distance** of each value from that mean (subtract the mean from each value, ignore minus signs)
- 3. Then find the **mean of those distances Example: the Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16**

Step 1: Find the **mean**:

$$\text{Mean} = \frac{3 + 6 + 6 + 7 + 8 + 11 + 15 + 16}{8} = \frac{72}{8} = 9$$

Step 2: Find the **distance** of each value from that mean:

Value	Distance from 9
3	6
6	3
6	3
7	2
8	1
11	2
15	6
16	7



Step 3. Find the **mean of those distances**:

$$\textbf{Mean Deviation} = \frac{6 + 3 + 3 + 2 + 1 + 2 + 6 + 7}{8} = \frac{30}{8} = 3.75$$

So, the **mean = 9**, and the **mean deviation = 3.75**

It tells us how far, on average, all values are from the middle.

In that example the values are, on average, 3.75 away from the middle. For **deviation** just think **distance** Formula

The formula is:

$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N}$$

Let's learn more about those symbols! Firstly:

- **μ** is the mean (in our example **μ = 9**)
- **x** is each value (such as 3 or 16)
- **N** is the number of values (in our example **N = 8**)

Standard Deviation

The Standard Deviation is a measure of how spread out numbers are. Its symbol is **σ** (the greek letter sigma)

The formula is easy: it is the **square root** of the **Variance.** So now you ask, "What is the Variance?"

Variance:

The Variance is defined as. The average of the squared differences from the Mean.

Examples of Standard Deviation:

This tutorial is about some examples of standard deviation using all methods which are discussed in the previous tutorial.

Example:

Calculate the standard deviation for the following sample data using all methods: 2, 4, 8, 6, 10, and 12.

Method-I: Actual Mean Method

Marks	f	X	*F(X)*	$(X-\bar{X})^2$	$f(X-X)^2$
1−3	40	2	80	4	160
3−5	30	4	120	0	0
5−7	20	6	120	4	80
7−9	10	8	80	16	160
Total	100		400		400

$\bar{X}=\sum f(X) / \sum f$ =400/ 100=4

Method-II: Taking assumed mean as 2

Marks	*f*	*X*	*D*=(*X*−2)	*fD*	*fD*2
1−3	40	2	0	0	0
3−5	30	4	2	60	120
5−7	20	6	4	80	320
7−9	10	8	6	60	160
Total	100			200	800

SKEWNESS

Lack of symmetry is called Skewness. If a distribution is not symmetrical then it is called skewed distribution. So, mean, median and mode are different in values and one tail becomes longer than other. The skewness may be positive or negative.

Positively skewed distribution:

If the frequency curve has longer tail to right the distribution is known as positively skewed distribution and ***Mean > Median > Mode***.

Negatively skewed distribution:

If the frequency curve has longer tail to left the distribution is known as negatively skewed distribution and ***Mean < Median < Mode***.

<u>Measure of Skewness</u>:

The difference between the mean and mode gives as absolute measure of skewness. If we divide this difference by standard deviation we obtain a relative measure of skewness known as coefficient and denoted by ***SK***.

Karl Pearson coefficient of Skewness

$$SK = Mean - Mode / S.D$$

Sometimes the mode is difficult to find. So we use another formula

$$SK = 3(Mean - Median)/S.D$$

Bowley_s coefficient of Skewness

$$SK = Q1 + Q3 - 2Median/Q3 - Q1$$

Kelly_s Measure of Skewness is one of several ways to measure skewness in a data distribution. Bowley_s skewness is based on the middle 50 percent of the observations in a data set. It leaves 25 percent of the observations in each tail of the distribution. Kelly suggested that leaving out fifty percent of data to calculate skewness was too extreme. He created a measure to find skewness with more data. Kelly_s measure is based on P_{90} (the 90th percentile) and P_{10} (the 10th percentile). Only twenty percent of observations (ten percent in each tail) are excluded from the measure**.**

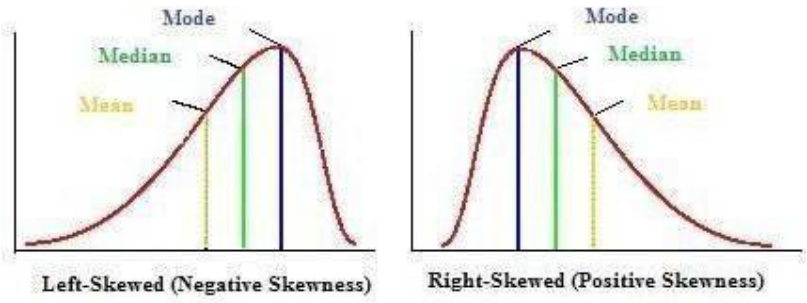
Kelly's Measure Formula.

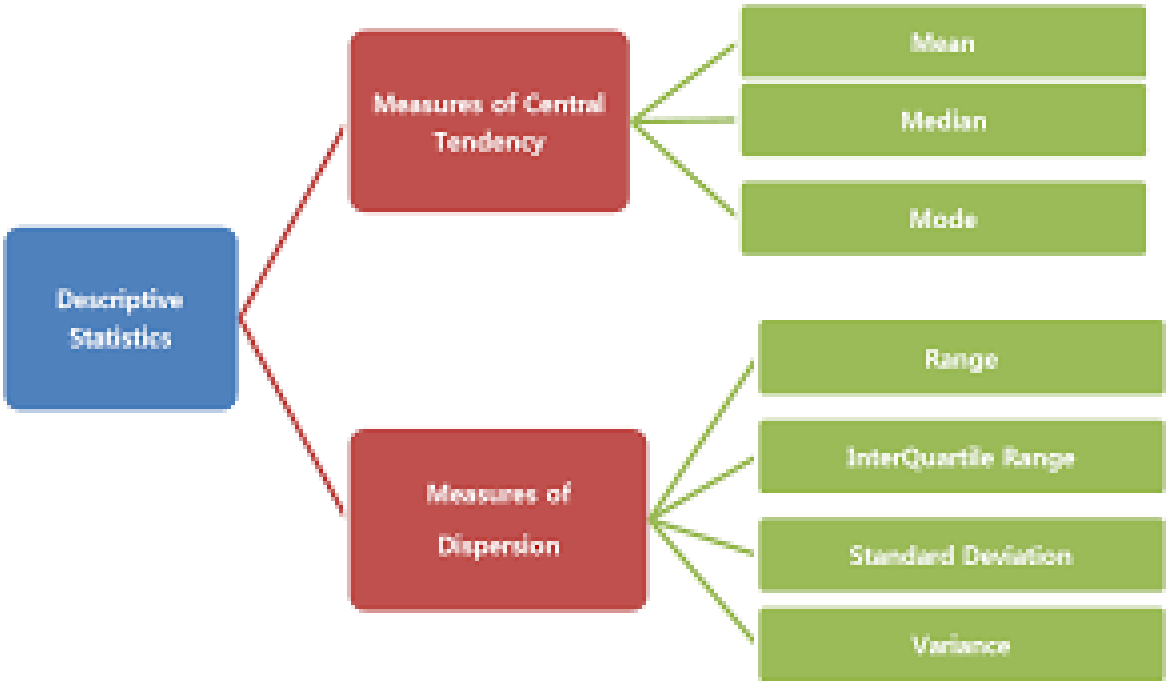
Kelley_s measure of skewness is given in terms of percentiles and deciles(D). Kelley_s absolute measure of skewness (S_k)is:

$S_k = P_{90} + P_{10} - 2*P_{50} = D_9 + D_1 - 2*D_5$.

Kelly_s Measure of Skewness gives you the same information about skewness as the other three types of skewness measures

A measure of skewness = 0 means that the distribution is symmetrical. A measure of skewness > 0 means a positive skewness.
A measure of skewness < means a negative skewness.




Descriptive Statistics
Measures of Central Tendency
Mean
Median
Mode
Measures of Dispersion
Range
InterQuartile Range
Standard Deviation
Variance

Theory of probability

CALCULATION AND CHANCE:

Most experimental searches for paranormal phenomena are statistical in nature. A subject repeatedly attempts a task with a known probability of success due to chance, and then the number of actual successes is compared to the chance expectation. If a subject scores consistently higher or lower than the chance expectation after a large number of attempts, one can calculate the probability of such a score due purely to chance, and then argue, if the chance probability is sufficiently small, that the results are evidence for the existence of some mechanism (precognition, telepathy, psycho kinesis, cheating, etc.) which allowed the subject to perform better than chance would seem to permit.

Suppose you ask a subject to guess, before it is flipped, whether a coin will land with heads or tails up. Assuming the coin is fair (has the same probability of heads and tails), the chance of guessing correctly is 50%, so you'd expect half the guesses to be correct and half to be wrong. So, if we ask the subject to guess heads or tails for each of 100 coin flips, we'd expect about 50 of the guesses to be correct. Suppose a new subject walks into the lab and manages to guess heads or tails correctly for 60 out of 100 tosses. Evidence of precognition or perhaps the subject's possesses a telekinetic power which causes the coin to land with the guessed face up? Well…no. In all likelihood, we've observed nothing more than good luck. The probability of 60 correct guesses out of 100 is about 2.8%, which means that if we do a large number of experiments flipping 100 coins, about every 35 experiments we can expect a score of 60 or better, purely due to chance.

But suppose this subject continues to guess about 60 right out of a hundred, so that after ten runs of 100 tosses—1000 tosses in all, the subject has made 600 correct guesses. The probability of *that* happening purely by chance is less than one in seven billion, so it's time to start thinking about explanations other than luck. Still, improbable things happen all the time: if you hit a golf ball, the odds it will land on a given blade of grass are millions to one, yet (unless it ends up in the lake or a sand trap) it is certain to land on *some* blades of grass.

Finally, suppose this "dream subject" continues to guess 60% of the flips correctly, observed by multiple video cameras, under conditions prescribed by skeptics and debunkers, using a coin provided and flipped by The Amazing Randi himself, with a final tally of 1200 correct guesses in 2000 flips. You'd have to try the 2000 flips more than 5×10^{18} times before you'd expect that result to occur by chance. If it takes a day to do 2000 guesses and coin flips, it would take more than 1.3×10^{16} years of 2000 flips per day before you'd expect to see 1200 correct guesses due to chance. That's more than a million times the age of the universe, so you'd better get started soon!

Claims of evidence for the paranormal are usually based upon statistics which diverge so far from the expectation due to chance that some other mechanism seems necessary to explain the experimental results. To interpret the results of our Retro Psycho Kinesis experiments, we'll be using the mathematics of probability and statistics, so it's worth spending some time explaining how we go about quantifying the consequences of chance.

PROBABILITY:

The word “Probability” or “Chance” is very commonly used in day-to-day conversation and generally people have a vague idea about its meaning. Example: We come across statements like “Probability it may rain tomorrow”.

The theory of Probability has its origin in the games of chance of chance related to Gambling games such as “Throwing a Die”, “Tossing Coin”, “Drawing cards from a well shuffled pack of cards” etc., ‘Galileo’ (1564 – 1642), an Italian mathematician, was the first man to attempt Quantitative measure of probability while dealing with some problems related to the theory of dice in gambling. Starting with games of chance, ‘Probability’ today has become one of the tools of statistics. In fact, statistics and probability are so fundamentally interrelated that it is difficult to discuss statistics without an understanding of the meaning of probability.

A Knowledge of Probability theory makes it possible to interpret statistical results, since many statistical procedures involve conclusions based on samples which are always affected by random variation, and it is by means of probability theory that we can express numerically the inevitable uncertainties in the resulting conclusions.

Probability theory is being applied in the 19^{th} century, required precise knowledge about the risk of loss in order to calculate premium. Within a few decades many learning centers were studying, probability as a tool for understanding social phenomena.

Today the concept of probability has assumed great importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision making research. In fact, probability has become a part of our everyday life. In personal and management decisions, we face uncertainty and use probability theory. In many instances we as concerned citizens, will have some knowledge about the possible outcomes of a decision.

Definition: (Probability):

The probability of a given event is an expression of likelihood or chance of occurrence of an event. A probability is a number which ranges from ‘0’ (zero) to ‘1’ (one) – “zero is for an event which cannot occur” and “one is for the event which occurs”. How the number is assigned would depend on the interpretation of the term ‘Probability’. There is no general agreement about its interpretation and many people associate probability and chance with nebulous and mystic ideas. However, broadly speaking, there are four different schools of thought on the concept of probability.

BASIC CONCEPTS:

1) Random experiment: An experiment is called a “Random experiment” if when conducted repeatedly under essentially homogenous conditions, the result is not unique but may be any one of the possible outcomes.

OR An experiment is said to be a random experiment, if it's out-come can't be predicted with certainty.
Example: If a coin is tossed, we can't say, whether head or tail will appear. So it is a random experiment.

2) Sample Space: The sent of all possible out-comes of an experiment is called the sample space. It is denoted by 'S' and its number of elements are n(s).

Example: In throwing a dice, the number that appears at top is any one of 1, 2, 3, 4, 5, 6. so here: S = {1, 2, 3, 4, 5, 6} and n(s) = 6
Similarly, in the case of a coin, S= {Head, Tail} or {H, T} and n(s) =2.
The elements of the sample space are called sample point or event point.

3) Event: Every subset of a sample space is an event. It is denoted by 'E'.
Example: In throwing a dice S={1,2,3,4,5,6}, the appearance of an event number will be the event E={2,4,6}. Clearly E is a sub set of S.

4) Simple event: An event, consisting of a single sample point is called a simple event.
Example: In throwing a dice, S={1,2,3,4,5,6}, so each of {1},{2},{3},{4},{5} and {6} are simple events.

5) Compound event: A subset of the sample space, which has more than on element is called a mixed event.
Example: In throwing a dice, the event of appearing of odd numbers is a compound event, because E = {1, 3, 5} which has '3' elements.

6) Equally likely events: Events are said to be equally likely, if we have no reason to believe that one is more likely to occur than the other.
Example: When a dice is thrown, all the six faces {1, 2, 3, 4, 5, and 6} are equally likely to come up.

7) Exhaustive events: The total number of possible outcomes of a random experiment is called the exhaustive cases for the experiment.
Example: (i) In toss of a single coin, we can get head (H) or Tail (T), Hence the exhaustive number of cases is 2 i.e., (H, T).
(ii) If two coins are tossed, the various possibilities are HH, HT, TH, TT, where 'HT' means head on the first coin and tail on the second coin and soon. Thus incase of tossing two coins, the exhaustive number of events are 4 i.e., 2^2.
(iii) Similarly, in a toss of three coins the possible number of outcomes is:
(H T) x (H T) x (H T) = (HHH, HTH, THH, TTH, HHT, HTT, THT, TTT)
Therefore, in case of toss of 3 coins the exhaustive number of cases are 8 i.e., $2^3 = 8$.

NOTE: In general, in a toss of 'n' coins (dice), the exhaustive number of cases is 2^n (6^n).

8) Mutually exclusive or disjoint event: If two or more events can't occur simultaneously, that is no two of them can occur together.
Example: When a coin is tossed, the event of occurrence of a head and the event of occurrence of a tail are mutually exclusive events.

9) Independent or mutually independent events: Two or more events are said to be independent if occurrence or non-occurrence of any of them does not affect the probability of occurrence or non-occurrence of the other event.
Example: When a coin is tossed twice, the event of occurrence of head in the first throw and the event of occurrence of head in the second throw are independent events.

10) Difference between mutually exclusive a mutually independent events: Mutually exclusiveness id used when the events are taken from the same experiment, where as independence is used when the events are taken from different experiments.

11) Favorable cases or Events: The number of outcomes of a random experiment which entail (result in) the happening of an event are termed as the cases favorable to the event.
Example: In toss of two coins, the number of cases favorable to the event 'exactly one head' is 2 i.e., HT, TH and for getting 'two heads' is one i.e., HH.

12) Complementary Events: Two events of a sample space whose intersection is Ø and whose union is the entire sample space are called "Complementary Events". Thus if E is an event of a sample space S, its complement is denoted by E^1 or $\bar{E}$ and $E \cap \bar{E} = \emptyset$, $E \cup \bar{E} = S$, $\bar{E} = S - E$.

APPROCHES TO PROBABILITY:

There are three approaches to probability. They are as follows.

3. CLASSICAL APPROACH
3. EMPIRICAL APPROACH
3. AXIOMATIC APPROACH

1. CLASSICAL or MATHEMATICAL or 'A PRIORI' APPROACH:

Definition: *If a random experiment in N exhaustive, mutually exclusive and equally likely outcomes (cases) out of which 'm' are favorable to the happening of an event A, then the probability of occurrence of A, usually denoted by P(A) and is given by:*

$$P(A) = \frac{\text{Favorable number of cases to A}}{\text{Exhaustive number of cases}}$$

This definition was given by 'James Bernouli' who was the first man to obtain a quantitative measure of uncertainty.

2. STATISTICAL OR EMPIRICAL PROBABILITY:

Definition: *If an experiment is performed repeatedly under essentially homogenous and identical conditions, then the limiting value of the ratio of the number of times the event occurs to the number of trials, as the number of trials becomes indefinitely large, is called the probability of happening of the event, it being assumed that the limit is finite and unique.*

Suppose that an event *A* occurs *m* times in *N* repetitions of a random experiment. Then the ratio *m/N* gives the relative frequency of the event *A* and it will not vary appreciably from one trial to another. In the limiting case when *N* becomes sufficiently large, it more or less settles to a number which is called the probability of *A*. Symbolically.

$$P(A) = \lim_{N \to \infty} (m/N)$$

3. AXIOMATIC PROBABILITY:

Definition: Given a sample space of a random experiment, the probability of the occurrence of any event *A* is defined as a set function *P (A)* satisfying the following axioms.

Axiom 1: P (A) is defined, is real and non-negative i.e.,
$P(A) \geq 0$ (Axiom of non-negativity)

Axiom 2: P (S) = 1 (Axiom of certainty)

Axiom 3: If $A_1, A_2, A_3 \ldots A_n$ is any finite or infinite sequence of disjoint events of S, then

$$P\left(\bigcup_{i=1} A_i\right) = \sum_{i=1} P(A_i) \quad \text{or} \quad P\left(\bigcup_{i=1} A_i\right) = \sum_{i=1} P(A_i)$$ (Axiom of Probability)

BASIC PROBLEMS ON PROBABILITY:

1) A uniform die is thrown at random. Find the probability that the number on it is:
i) 5 ii) Greater than 4 iii) Even

Solution: Since the dice can fall with any one of the faces 1, 2, 3, 4, 5 and 6. Therefore the exhaustive number of cases N = 6.

i) The number of cases favorable to the event 'A' getting '5' is only 1 i.e., m = 1.
Therefore, the required probability is P (A) = m/N = 1/6.

ii) The number of cases favorable to the event 'B' of getting a number greater than 4 is m = 2 i.e., 5 and 6.
Therefore, the required probability is P (B) = m/N = 2/6 = 1/3.

iii) Favorable cases to the event 'C' for getting an even number are 2, 4 and 6 i.e., m = 3.
Therefore, the required probability is P (C) = m/N = 3/6 = 1/2.

2) In a single throw with two uniform dice find the probability of throwing:

i) Five ii) Eight

Solution: Exhaustive number of cases in a single throw with two dice is $6^2 = 36$.

i) Sum of '5' can be obtained on the two dice in the following mutually exclusive ways: (1, 4), (2, 3), (3, 2), (4, 1) i.e., 4 cases in all where the first and second number in the bracket ( ) refer to the numbers on the first and second dice respectively.

Therefore, the required probability = m/N = 4/36 = 1/9.

ii) The cases favorable to the event of getting sum of 8 on two dice are:

(2, 6), (3, 5), (4, 4), (5, 3), (6, 2) i.e., 5 distinct cases in all.

Therefore, the required probability = m/N = 5/36.

3) What is the chance that a non – leap year should have fifty-three Sundays?

Solution: A non – leap year consists of 365 days i.e., 52 full weeks and one over day. A non- leap year will consist of 53 Sundays if this over day is a Sunday. This over day can be anyone of the possible outcomes: Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday i.e., 7 outcomes in all. Of these, the number of ways favorable to the required event is if the over day being Sunday i.e., m = 1. The exhaustive number of cases N = 7.

Therefore, the required probability = 1/7.

4) Four cards can be drawn at random from a pack of 52 cards. Find the probability that:

i) They are a king, a queen, a jack, and an ace.

ii) Two are kings and two are aces.

iii) All are diamonds.

iv) Two are red and two are black.
v) There is one card from each suit.
vi) There are two cards of clubs and two cards of diamonds.

Solution: Four cards can be drawn from a well shuffled pack of 52 cards in $52C_4$ ways, which gives the exhaustive number of cases.

(i) 1 king can be drawn out of the 4 kings in $4c_1 = 4$ ways. Similarly, 1 queen, 1 jack and an ace can be drawn in $4c_1 = 4$ ways. Since any one of the ways of drawing a king can be associated with any one of the ways of drawing a queen, a jack, and an ace, the favorable number of cases are $4c_1 \times 4c_1 \times 4c_1 \times 4c_1$.

Hence, required probability = $\frac{4c_1 \times 4c_1 \times 4c_1 \times 4c_1}{52c_4} = \frac{256}{52c_4}$

(ii) Required probability = $\frac{4c_2 \times 4c_2}{52c_4}$

(iii) Since 4 cards can be drawn out of 13 cards (since there are 13 cards of diamond in a pack of cards) in $13c_4$ ways,

Therefore, the required probability = $\frac{13c_4}{52c_4}$

(iv) Since there are 26 red cards (of diamonds and hearts) and 26 black cards (of spades and clubs) in a pack of cards.

Therefore, the required probability = $\frac{26c_2 \times 26c_2}{52c_4}$

(v) Since, in a pack of cards there are 13 cards of each suit,

Then the required probability = $\frac{13c_1 \times 13c_1 \times 13c_1 \times 13c_1}{52c_4}$

(vi) The required probability = $\frac{13c_2 \times 13c_2}{52c_4}$

5) At a car park there are 100 vehicles, 60 of which are cars, 30 are vans and the remainder are Lorries. If every vehicle is equally likely to leave, find the probability of:

a) Van leaving first.

b) Lorry leaving first.

c) Car leaving second if either a lorry or van had left first.

Solution:
a) Let *S* be the sample space and *A* be the event of a van leaving first.
n(*S*) = 100 , n(*A*) = 30

Probability of a van leaving first:

$$P(A) = \frac{30}{100} = \frac{3}{10}$$

b) Let *B* be the event of a lorry leaving first. n(*B*) = 100 – 60 – 30 = 10

Probability of a lorry leaving first:

$$P(B) = \frac{10}{100} = \frac{1}{10}$$

c) If either a lorry or van had left first, then there would be 99 vehicles remaining, 60 of which are cars. Let *T* be the sample space and *C* be the event of a car leaving.
n(*T*) = 99 n(*C*) = 60

Probability of a car leaving after a lorry or van has left:

$$P(C) = \frac{60}{99} = \frac{20}{33}$$

6) A survey was taken on 30 classes at a school to find the total number of left-handed students in each class. The table below shows the results:

No. of left-handed students	**0**	**1**	**2**	**3**	**4**	**5**
Frequency (no. of classes)	**1**	**2**	**5**	**12**	**8**	**2**

A class was selected at random.
a) Find the probability that the class has 2 left-handed students.
b) What is the probability that the class has at least 3 left-handed students?
c) Given that the total number of students in the 30 classes is 960, find the probability that a student randomly chosen from these 30 classes is left-handed.

Solution:
a) Let *S* be the sample space and *A* be the event of a class having 2 left-handed students.
n(*S*) = 30 n(*A*) = 5

$$P(A) = \frac{5}{30} = \frac{}{6}$$

b) Let *B* be the event of a class having at least 3 left-handed students.
n(*B*) = 12 + 8 + 2 = 22

$$P(B) = \frac{22}{30} = \frac{11}{15}$$

c) First find the total number of left-handed students:

No. of left-handed students, *x*	0	1	2	3	4	5
Frequency, *f* **(no. of classes)**	1	2	5	12	8	2
fx	0	2	10	36	32	10

Total no. of left-handed students = 2 + 10 + 36 + 32 + 10 = 90

Here, the sample space is the total number of students in the 30 classes, which was given as 960. Let *T* be the sample space and *C* be the event that a student is left-handed.
n(*T*) = 960 n(*C*) = 90

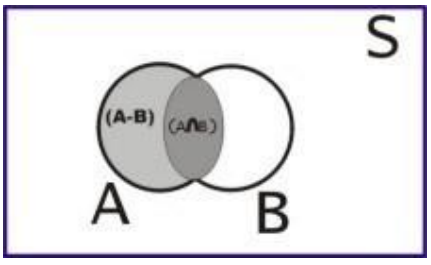
$$P(C) = \frac{90}{960} = \frac{3}{32}$$

ADDITION THEOREM OF PROBABILITY:

Theorem: If 'A' and 'B' by any two events, then the probability of occurrence of at least one of the events 'A' and 'B' is given by:

Proof: P (A or B) = P (A) + P (B) – P (A and B)

P (A ∪ B) = P (A) + P (B) – P (A ∩ B)



From set theory, we have:

n (A ∪ B) = n(A) + n(B) – n(A ∩ B)

Dividing both sides by n(S):

$n(A \cup B) / n(S) = n(A) / n(S) + n(B) / n(S) - n(A \cap B) / n(S)$

Or $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Corollary: If 'A' and 'B' are mutually exclusive events,

Then $P(A \cap B) = 0$. [As we have proved]

In this case:

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

Addition theorem for '3' events 'A', 'B' and 'C':

$P(A \cup B \cup C) = P(A) + P(B) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

Proof: $P(A \cup B \cup C) = P[(A \cup B) \cup C]$

$= P(A \cup b) + P(C) - P[(A \cup B) \cap C]$ [By addition theorem for two events]

$= P(A \cup B) - P(C) - [P(A \cap C) + P(B \cap C) - P(A \cap C \cap B \cap C)]$

$= P(A) + P(B) - P(A \cap B) + P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

Corollary: If 'A', 'B' and 'C' are mutually exclusive events, then $P(A \cap B) = 0$, $P(B \cap C) = 0$, $P(A \cap C) = 0$ and $P(A \cap B \cap C) = 0$.

In this case:

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

PROBLEMS

Problems based on addition theorem of probability:

1) The probability that a contractor will get a contract is '2/3' and the probability that he will get on other contract is 5/9. If the probability of getting at least one contract is 4/5, what is the probability that he will get both the contracts?

Sol.: Here $P(A) = 2/3$, $P(B) = 5/9$

$P(A \cup B) = 4/5$, $P(A \cap B) = ?$

By addition theorem of Probability:

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

$= 4/5 = 2/3 + 5/9 - P(A \cap B)$

Or $4/5 = 11/9 - P(A \cap B)$

Or $P(A \cap B) = 11/9 - 4/5 = (55-36) / 45$

$P(A \cap B) = 19/45$

2) Two cards are drawn at random. Find the probability that either the cards are of red color or they are queen.

Sol.: Let S = Sample – space.

A = the event that the two cards drawn are red.

B = the event that the two cards drawn are queen.

⇨ $A \cap B$ = the event that the two cards drawn are queen of red color.

⇨ $n(S) = {}^{52}C_2$, $n(A) = {}^{26}C_2$, $n(B) = {}^{4}C_2$

$n(A \cap B) = {}^{2}C_2$

⇨ $P(A) = n(A) / n(S) = {}^{26}C_2 / {}^{52}C_2$,

$P(B) = n(B) / n(S) = {}^{4}C_2 / {}^{52}C_2$

$P(A \cap B) = n(A \cap B) / n(S) = {}^{2}C_2 / {}^{52}C_2$

$P(A \cup B) = ?$

We have $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$= {}^{26}C_2 / {}^{52}C_2 + {}^{4}C_2 / {}^{52}C_2 - {}^{2}C_2 / {}^{52}C_2$

$= ({}^{26}C_2 + {}^{4}C_2 - {}^{2}C_2) / {}^{52}C_2$

$= (13X25+2X3-1) / (26X51) \Rightarrow P(A \cup B) = 55/221$

3) A bag contains '6' white and '4' red balls. Two balls are drawn at random. What is the chance, they will be of the same color?

Sol) Let S = Sample space

A = the event of drawing ‘2’ white balls.

B = the event of drawing ‘2’ red balls.

$A \cup B$ = the event of drawing 2 white balls or 2 red balls.

I.e. the event of drawing ‘2’ balls of same color.

⇨ $n(S) = {}^{10}C_2 = 10! / (2! \times 8!) = 45$

$n(A) = {}^{6}C_2 = 6! / ((2! \times 4!) = (6 \times 5) / 2 = 15$

$n(B) = {}^{4}C_2 = 4! / (2! \times 2!) = (4 \times 3) / 2 = 6$

$P(A) = n(A) / n(S) = 15/45 = 1/3$

$P(B) = n(B) / n(S) = 6/45 = 2/15$

⇨ $P(A \cup B) = P(A) + P(B)$

⇨ $1/3 + 2/15 = (5+2) / 15$

$P(A \cup B) = 7/15$

4) For a post three persons ‘A’, ‘B’ and ‘C’ appear in the interview. The probability of ‘A’ being selected is twice that of ‘B’ and the probability of ‘B’ being selected is thrice that of ‘C’, what are the individual probabilities of A, B, C being selected?

Sol) Let ‘E_1’, ‘E_2’, ‘E_3’ be the events of selections of A, B, and C respectively.

Let $P(E_3) = x$

⇨ $P(E_2) = 3. P(E_3) = 3x$

And $P(E_1) = 2P(E_2) = 2 \times 3x = 6x$

As there are only ‘3’ candidates ‘A’, ‘B’ and ‘C’ we have to select at least one of the candidates A or B or C, surely.

⇨ $P(E_1 \cup E_2 \cup E_3) = 1$

And E_1, E_2, E_3 are mutually exclusive.

⇨ $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3)$

$1 = 6x + 3x + x$ ⇨ $10x – 1$ or $x = 1/10$

⇨ $P(E_3) = 1/10$, $P(E_2) = 3/10$ and $P(E_1) = 6/10 = 3/5$

5) A bag contains 4 green, 6 black and 7 white balls. A ball is drawn at random. What is the probability that it is either a green or a black ball?

Sol) Let S is the sample space associated with the drawing of a ball from a bag containing 4 green, 6 black and 7 white balls.

⇨ $n(S) = 17c_1 = 17$

Let E_1 denote the event of drawing a green ball and E_2 be the event of drawing a black ball.

⇨ $n(E_1) = 4c_1 = 4$, $n(E_2) = 6c_1 = 6$

Then the required probabilities of events E_1, E_2 are as follows:

⇨ $P(E_1) = n(E_1) / n(S) = 4/17$

⇨ $P(E_2) = n(E_2) / n(S) = 6/17$

Since, E_1, E_2 are two mutually exclusive events i.e. $E_1 \cap E_2 = \emptyset$

⇨ $P(E_1 \cap E_2) = P(\emptyset) = 0$ (By Theorem)

⇨ $P(E_1 \cap E_2) = 0$

By the Addition theorem of probability for two events E_1, E_2 is:

$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

Therefore, the probability of drawing either a green ball or a black ball = $P(E_1 \cup E_2)$

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$ [since, $P(E_1 \cap E_2) = 0$]

⇨ $P(E_1 \cup E_2) = 4/17 + 6/17$

⇨ $P(E_1 \cup E_2) = 10/17$

Hence, the probability of drawing either a green or black ball is 10/17.

6) Three students A, B, C are in running race. A and B have the same probability of winning and each is twice as likely to win as C. Find the probability that B or C wins.

Sol) Let A, B, C are three events (Students) in a running race.

$A \cup B \cup C = S$ = Sample Space of race. From the given problem, we have:

$P(A) = P(B)$, $P(A) = 2\,P(C), P(B) = 2\,P(C)$

Since the three events are mutually exclusive events, we have:

$P(A \cap B) = P(B \cap C) = P(A \cap C) = P(A \cap B \cap C) = 0$.

Also, we know that, $P(A) + P(B) + P(C) = 1$

⇨ $2\,P(C) + 2\,P(C) + P(C) = 1$

⇨ $5\,P(C) = 1$

⇨ $P(C) = 1/5$

Therefore, the individual probabilities of the three students (events) are as follows:

$P(A) = 2/5, \quad P(B) = 2/5, \quad P(C) = 1/5$

Then, the probability that B or C wins = $P(B \cup C) = P(B) + P(C) – P(B \cap C)$

- $P(B \cup C) = 2/5 + 1/5 – 0$ [since, B & C are mutually exclusive events]
- $P(B \cup C) = 3/5$.

Hence, the probability that student B or student C wins is 3/5.

7) A card is drawn from a well shuffled pack of cards. What is the probability that it is either a spade or an ace?

Sol) Let S is the sample space of all the simple events.

Therefore, $n(S) = 52c_1 = 52$

Let A denote the event of getting a spade and B denotes the event of getting an ace. $n(A) = 13c_1 = 13$, $n(B) = 4c_1 = 4$, $n(A \cap B) = 1c_1 = 1$

Then, $A \cup B$ = the event of getting a spade or an ace.

$A \cap B$ = the event of getting a spade and an ace.

Therefore, the probabilities of these events are as follows:

$P(A) = 13/52, \quad P(B) = 4/52, \quad P(A \cap B) = 1/52$

By the Addition theorem of probability for two events A & B is:

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

- $P(A \cup B) = 13/52 + 4/52 – 1/52$
- $P(A \cup B) = (13 + 4 – 1) / 52$
- $P(A \cup B) = 16/52 = 4/13$
- $P(A \cup B) = 4/13$

Hence, the probability of getting either a spade or an ace is 4/13.

8) Find the probability of getting a sum of 10 if we throw two dice?

Sol)

When two dice are thrown, number of possible outcomes = $n(S) = 6^2 = 36$

Let E is the event of getting the sum as 10.

Favorable outcomes are (4, 6) (5, 5) (6, 4) i.e. $n(E) = 3$

Thus, the probability of the event getting the sum as 10 is:

$P(E) = n(E) / n(S) = 3/36 = 1/12$

Hence, the probability of the event getting the sum as 10 is 1/12.

9) From a city 3 news papers A, B, C are being published. A is read by 20%, B is read by 16%, C is read by 14%, both A & B are read by 8%, both A & C are read by 5%, both B & C are read by 4% and all three A, B, C are read by 2%. What is the percentage of the population that read at least one paper?

Sol)

Given, the probabilities of A, B, C are:

P(A) = 20/100, P(B) = 16/100, P(C) = 14/100, P(A ∩ B) = 8/100, P(B ∩ C) = 4/100,

P(A ∩ C) = 5/100 and P(A ∩ B ∩ C) = 2/100.

By the Addition theorem of probability for three events A, B, C is as follows:

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

⇨ $P(A \cup B \cup C) = 20/100 + 16/100 + 14/100 - 8/100 - 4/100 - 5/100 + 2/100$

⇨ $P(A \cup B \cup C) = (20 + 16 + 14 - 8 - 4 - 5 + 2) / 100$

⇨ $P(A \cup B \cup C) = 35/100$

⇨ $P(A \cup B \cup C) = 35\%$

CONDITIONAL EVENT:

Definition: If E_1, E_2 are two events of a sample space S and if E_2 occurs after the occurrence of E_1, then the event of occurrence of E_2 after the event E_1 is called "Conditional Event of event E_2 given E_1". It is denoted by E_2 / E_1. Similarly, we define E_1 / E_2.

Examples:

1. Two coins are tossed. The event of getting two tails given that there is at least one tail is a conditional event.

2. Two unbiased dice are thrown. If the sum of the numbers thrown on them is 7, the event of getting 1 on anyone of them is a conditional event.

3. A die is thrown 3 times. The event of getting the sum of the numbers thrown is 15 when it is known that the first throw was a 5 is a conditional event.

CONDITIONAL PROBABILITY:

Definition: If E_1 and E_2 are two events in a sample space S and $P(E_1) \neq 0$, then the probability of E_2, after the event E_1 has occurred, is called the "Conditional Probability of the event E_2 given E_1 and is denoted by $P(E_2 / E_1)$ and we define as:

$$P(E_2 / E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} = \frac{n(E_1 \cap E_2) / n(S)}{n(E_1) / n(S)} = n(E_1 \cap E_2) / n(E_1)$$

Similarly, we define $P(E_1 / E_2) = \dfrac{P(E_2 \cap E_1)}{P(E_2)}$

Some definitions based on Conditional probability:

1) Compound Event: When two or more events occur in conjunction with each other, their joint occurrence is called "Compound Event".

Examples:

i. If 2 balls are drawn from a bag containing 4 green, 6 black and 7 white balls, the event of drawing 2 green balls or 2 white balls is a Compound event.
ii. When a die and a coin are tossed the event of getting utmost 4 on the die and head on the coin is a compound event and separately they are independent events.
NOTE: Multiplication theorem of probability is also called theorem of compound probability.

2) Independent Events: If the occurrence of the event E_2 is not affected by the occurrence or non-occurrence of the event E_1, then the event E_2 is said to be independent of E_1 and $P(E_2 / E_1) = P(E_2)$, similarly, we define, $P(E_1 / E_2) = P(E_1)$.

3) Mutually Independent or Simply Independent: If $P(E_1) \neq 0$, $P(E_2) \neq 0$ and E_2 is independent of E_1, then E_1 is independent of E_2. In this case we say that E_1, E_2 are "Mutually independent or simply independent events".

4) Dependent Events: If the occurrence of the event E_2 is affected by the occurrence of E_1, then the events E_1, E_2 are dependent and $P(E_2 / E_1) \neq P(E_2)$.

MULTIPLICATION THEOREM OF PROBABILITY:

Theorem: In a random experiment, if E_1, E_2 are two events such that $P(E_1) \neq 0$, $P(E_2) \neq 0$, then

$$\mathbf{P(E_1 \cap E_2) = P(E_2) \,.\, P(E_1 / E_2)}$$

$$\mathbf{P(E_2 \cap E_1) = P(E_1) \,.\, P(E_2 / E_1)}$$

Proof: Let S be the sample space associated with the random experiment. Let E_1, E_2 be two events of S such that $P(E_1) \neq 0$, $P(E_2) \neq 0$. Since, $P(E_1) \neq 0$, by the definition of conditional probability of E_2 given E_1,

$$P(E_2 / E_1) = P(E_1 \cap E_2) / P(E_1)$$

$$\Rightarrow \quad P(E_1 \cap E_2) = P(E_1) \,.\, P(E_2 / E_1)$$

Since, $P(E_2) \neq 0$, we have, $P(E_1 / E_2) = P(E_2 \cap E_1) / P(E_2)$

$\Rightarrow$ $P(E_2 \cap E_1) = P(E_2) \cdot P(E_1 / E_2)$

NOTE: Multiplication theorem can be extended to three events i.e. E_1, E_2, & E_3 as:

$P(E_1 \cap E_2 \cap E_3) = P(E_1) \cdot P(E_1 / E_2) \cdot P(E_3 / (E_1 \cap E_2))$

This result can be extended to 4 or more events.

PROBLEMS ON MULTIPLICATION THEOREM:

1) A bag contains 8 red balls and 6 blue balls. Two drawing of each 2 balls are made. Find the probability that the first drawing gives two red balls and second drawing gives 2 blue balls, if the balls drawn are replaced before the second draw?

Sol) Let E_1 is the event of drawing 2 red balls in the first draw from the bag containing 8 red and 6 blue balls.

i.e. $n(E_1) = 8c_2 = 28$, $n(S) = 14c_2 = 91$

Therefore, $P(E_1) = 8c_2 / 14c_2 = 28/91$

Let E_2 is the event of drawing 2 blue balls in the second draw from the bag.

i.e. $n(E_1) = 6c_2 = 15$, $n(S) = 14c_2 = 91$

Therefore, $P(E_1) = 15/91$

Now, $E_1 \cap E_2$ = Event of drawing 2 red balls in the first draw and another drawing of 2 blue balls in the second draw after the balls are replaced.

Also, E_1, E_2 are independent.

Therefore, by Multiplication theorem of 2 independent events, we have:

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$

$\Rightarrow$ $P(E_1 \cap E_2) = (28/91) \times (15/91)$

$\Rightarrow$ $P(E_1 \cap E_2) = 420/8281$

$\Rightarrow$ $P(E_1 \cap E_2) = 60/1183$

Hence, the probability of drawing 2 red balls in the first draw and another drawing of 2 blue balls in the second draw after the balls are replaced is 60/1183.

2) Three students are chosen at random from a class consisting of 12 boys and 4 girls. Find the probability for three students chosen one after another in succession to be boys.

Solution) Let A, B, C be the three events of choosing a boy each time in succession.

Therefore, probability for 3 students chosen one after another in succession to be boys.

Hence, P(A) = 12/16, P(B) = 11/15, P(C) = 10/14

By multiplication theorem, we have: $P(A \cap B \cap C) = P(A) \cdot P(B/A) \cdot P[C/(A \cap B)]$

Since, A, B, C are independent events, we have: $P(A \cap B \cap C) = P(A).P(B).P(C)$

Therefore, $P(A \cap B \cap C) = (12/16) \times (11/15) \times (10/14)$

⇨ $P(A \cap B \cap C) = 1320/3360$

⇨ $P(A \cap B \cap C) = 66/168 = 11/28$

Hence, the probability for three students chosen one after another in succession to boys is 11/28.

3) A class has 10 boys and 5 girls. Three students are selected at random one after another. Find the probability that (i) first two are boys and third is girl (ii) first and third are of same sex and the second is of opposite sex.

Solution) The total number of students in a class = 10+5 = 15

(i) The probability that first two are boys and the third is girl is:

$P(E_1 \cap E_2 \cap E_3) = P(E_1) \cdot P(E_2) \cdot P(E_3)$ → (1)

Where, E_1 = Event of selecting the first student as boy

E_2 = Event of selecting the second student as boy

E_3 = Event of selecting the third student as a girl.

Therefore, from the given data, we have:

$P(E_1) = 10/15$, $P(E_2) = 9/14$, $P(E_3) = 5/13$

From eq(1), we have: $P(E_1 \cap E_2 \cap E_3) = P(E_1) \cdot P(E_2) \cdot P(E_3)$

⇨ $P(E_1 \cap E_2 \cap E_3) = (10/15) \cdot (9/14) \cdot (5/13)$

⇨ $P(E_1 \cap E_2 \cap E_3) = 15/91$

Hence, the probability of selecting first two students as boys and the third student as a girl is 15/91.

(ii) Suppose, the probability that the first and third are boys and second is a girl.

Let, E_1 = Event of selecting the first student as boy.

E_2 = Event of selecting the second student as girl.

E_3 = Event of selecting the third student as a boy.

Let A = the probability of the event that first and third students are boys and second student is a girl.

Therefore, $P(A) = P(E_1 \cap E_2 \cap E_3) = P(E_1) \cdot P(E_2) \cdot P(E_3)$

Where, $P(E_1) = 10/15$, $P(E_2) = 5/14$, $P(E_3) = 9/13$

⇨ $P(A) = (10/15) . (5/14) . (9/13)$

⇨ $P(A) = 15/91$

Similarly, suppose that, the probability that the first and third students are girls and second student is a boy.

Let, E_1 = Event of selecting the first student as girl.

E_2 = Event of selecting the second student as boy.

E_3 = Event of selecting the third student as a girl.

Let B = the probability of the event that first and third students are girls and second student is a boy.

Therefore, $P(B) = P(E_1 \cap E_2 \cap E_3) = P(E_1) . P(E_2) . P(E_3)$

Where, $P(E_1) = 5/15$, $P(E_2) = 10/14$, $P(E_3) = 4/13$

⇨ $P(B) = (5/15) . (10/14) . (4/13)$

⇨ $P(B) = 20/273$

Hence required probability = P(A) + P(B)

⇨ $(15/91) + (20/273)$

⇨ $65/273$

Hence, the probability of selecting first and third students are of same sex and the second student is of opposite sex is 65/273.

4) Two airplanes bomb a target in succession. The probability of each correctly scoring a hit is 0.3 and 0.2 respectively. The second will bomb only if the first misses the target. Find the probability that (i) target is hit (ii) both fails to score hits.

Solution) Let A = the event of first plane hitting the target.

B = the event of second plane hitting the target.

The probability of first plane hitting the target = P(A) = 0.3

The probability of second plane hitting the target = P(B) = 0.2

The probability that the first plane fails to hit the target = $P(A^1) = 1–P(A) = 0.7$

The probability that the second plane fails to hit the target = $P(B^1) = 1-P(B) = 0.8$

(i) P(target is hit) = P[(A hits) or (A fails and B hits)]

⇨ $P[A \cup (A^1 \cap B)]$

⇨ $P(A) + P(A^1 \cap B)$ (Since, by Addition theorem)

⇨ $P(A) + P(A^1) . P(B)$ (Since, by Multiplication theorem)

⇨ 0.3 + (0.7) x (0.2)

⇨ 0.3 + 0.14 ⇨ 0.44

Therefore, the probability that both airplanes hit the target is “0.44”.

(ii) P(both fails to hit) = P(A fails and B fails)

⇨ $P(A^1 \cap B^1)$

⇨ $P(A^1) . P(B^1)$

⇨ (0.7) x (0.8) ⇨ 0.56

Therefore, the probability that both airplanes fails to hit the target is “0.56”.

5) Determine (1) P(B/A) (2) $P(A/B^1)$, if A and B are two events with their probabilities P(A) = 1/3, P(B) = 1/4, $P(A \cup B) = 1/2$.

Solution) Given that, P(A) = 1/3, P(B) = 1/4, $P(A \cup B) = 1/2$

From the Addition theorem of probability, we have:

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

⇨ $P(A \cap B) = P(A) + P(B) - P(A \cup B)$ ⇨ $P(A \cap B) = 1/3 + 1/4 - 1/2$

⇨ $P(A \cap B) = 1/12$

(i) $P(B/A) = [P(A \cap B)] / P(A)$

⇨ $P(B/A) = (1/12) / (1/3) = 1/4$

Therefore, P(B/A) = 1/4.

(ii) $P(A/B^1) = [P(A \cap B^1)] / P(B^1)$, $P(B^1) = 1-P(B) = 1-1/4 = 3/4$

⇨ $P(A/B^1) = [P(A) - P(A \cap B)] / P(B^1)$

⇨ $P(A/B^1) = [1/3 - 1/12] / (3/4)$ ⇨ $P(A/B^1) = 1/3$.

Therefore, $P(A/B^1) = 1/3$.

6) In a certain town 40% have brown hair, 25% have brown eyes and 15% have both brown hair and brown eyes. A person is selected at random from the town.

(i) If he has brown hair, what is the probability that he has brown eyes also?

(ii) If he has brown eyes, determine the probability that he does not have brown hair?

Solution) Given,

n(S) = 100, and let A = the set of people who have brown hair.

Then, n(A) = 40, Therefore, P(A) = n(A) / n(S) = 40/100

Let B = the set of people who have brown eyes.

Then, n(B) = 25, Therefore, P(B) = n(B) / n(S) = 25/100

Let, A ∩ B = the set of people who have both brown hair and brown eyes.

Then, n(A ∩ B) = 15, Therefore, P(A ∩ B) = 15/100

(i) If the selected person has brown hair, then the probability that he has brown eyes also is P(B/A).

Therefore, P(B/A) = [P(A ∩ B)] / P(A)

⇨ P(B/A) = (15/100) / (40/100)

⇨ P(B/A) = 3/8

Hence, if the selected person has brown hair, then the probability that he has brown eyes also is 3/8.

(ii) If the selected person has brown eyes, then the probability that he does not have brown hair is $P(A^1/B)$.

Therefore, $P(A^1/B) = [P(A^1 \cap B)] / P(B) = [P(B) – P(A \cap B)] / P(B)$

⇨ $P(A^1/B) = [(25/100) – (15/100)] / (25/100)$

⇨ $P(A^1/B) = 2/5$

Hence, the selected person has brown eyes, then the probability that he does not have brown hair is 2/5.

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

Solution)

Suppose, E_1 = the event that the marble is drawn from box A and is red.

Therefore, $P(E_1) = (1/2) . (5/8) = 5/16$

E_2 = the event that the marble is from box B and is red.

Therefore, $P(E_2) = (1/2) . (2/8) = 1/8$

Hence, the probability that both the marbles are red is $P(E_1 \cap E_2)$.

i.e. $P(E_1 \cap E_2) = P(E_1) . P(E_2) = (5/16) . (1/8) = 5/128$.

Let E_3 = the event that the marble drawn from box A and is white.

Therefore, $P(E_3) = (1/2) . (3/8) = 3/16$

Let E_4 = the event that the marble drawn is from box B and is white.

Therefore, $P(E_4) = (1/2) . (6/8) = 3/8$

Hence, the probability that both the marbles are white is $P(E_3 \cap E_4)$.

i.e. $P(E_3 \cap E_4) = P(E_3) . P(E_4) = (3/16) . (3/8) = 9/128$.

The probability that the marbles are of same color = $P(E_1 \cap E_2) + P(E_3 \cap E_4)$

⇨ (5/128) + (9/128)

⇨ 14/128

⇨ 7/64

Hence, if a marble is drawn from each box, then the probability that they are of the same color is 7/64.

8) Two marbles are drawn in succession from a box containing 10 red, 30 white, 20 blue and 15 orange marbles, with replacement being made after each drawing. Find the probability that (i) both are white (ii) first is red and second is white

Solution) Given, the total number of marbles in the box = 75

(i) Let E_1 = Event of the first drawn marble is white.

Therefore, $P(E_1) = 30/75$

Let E_2 = Event of second drawn marble is also white

Therefore, $P(E_2) = 30/75$

The probability that both marbles are white (with replacement) is:

$P(E_1 \cap E_2) = P(E_1) . P(E_2) = (30/75) . (30/75)$ [Since, E_1, E_2 are Independent]

⇨ $P(E_1 \cap E_2) = 4/25$

Hence, the probability that both the marbles are white is 4/25.

(ii) Let E_1 = the event that the first drawn marble is red.

Therefore, $P(E_1) = 10/75 = 2/15$

Let E_2 = the event that second drawn is white.

Therefore, $P(E_2) = 30/75 = 2/5$

The probability that the first marble is red and the second marble is white is:

$P(E_1 \cap E_2) = P(E_1) . P(E_2) = (2/15) . (2/5)$ [Since, E_1, E_2 are Independent]

⇨ $P(E_1 \cap E_2) = 4/75$

Hence, the probability that first is red and second is white is 4/75.

9) Three boxes, practically indistinguishable in appearance have two drawers each. Box 1 contains a gold coin in first and silver coin in the other drawer, Box 2 contains a gold coin in each drawer and Box 3 contains a silver coin in each drawer. One box is chosen at random and one of its drawers is opened at random and a gold coin is found. What is the probability that the other drawer contains a coin of silver?

Solution) Let E_i denotes the event that the box is chosen; for i = 1, 2, 3

Therefore, $P(E_i) = 1/3$; for i = 1, 2, 3

Let A denotes the event that the gold coin is chosen. Then,

$P(A/E_i)$ = Probability that a gold coin is chosen from the box i =1, 2, 3

Therefore, $P(A/E_1) = 1/2$ (since, the total no. of coins in box 1 is 2)

$P(A/E_2) = 2/2 = 1$ (there are two gold coins in box 2)

$P(A/E_3) = 0/2 = 0$ (there is no gold coin in box 3)

(i) The probability that the drawn coin is gold

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

$= (1/3) . (1/2) + (1/3) .1 + (1/3) .(0)$

$= (1/6) + (1/3) = 1/2$

Therefore, $P(A) = 1/2$

(ii) The probability that the drawn coin is silver is $P(B) = 1 – P(A)$

i.e. $P(B) = 1 – 1/2 = 1/2$.

10) The probabilities that students A, B, C, D solve a problem are 1/3, 2/5, 1/5 and 1/4 respectively. What is the probability that the problem is solved?

Solution) Given the probability of A, B, C, D solving the problem is:

$P(A) = 1/3$, $P(B) = 2/5$, $P(C) = 1/5$, $P(D) = 1/4$

The probability that the problem is not solved by A, B, C, and D is:

$P(A^1) = 2/3$, $P(B^1) = 3/5$, $P(C^1) = 4/5$ and $P(D^1) = 3/4$

The probability that the problem is not solved when A, B, C, D try together (independently)

$P(A^1 \cap B^1 \cap C^1 \cap D^1) = P(A^1) . P(B^1) . P(C^1) . P(D^1)$

⇨ $P(A^1 \cap B^1 \cap C^1 \cap D^1) = 2/3 . 3/5 . 4/5 . 3/4 = 6/25$

The probability that the problem is solved if they try independently is:

$P(A \cup B \cup C \cup D) = 1 - P(A^1 \cap B^1 \cap C^1 \cap D^1)$

⇨ $P(A \cup B \cup C \cup D) = 1 – 6/25$

⇨ $P(A \cup B \cup C \cup D) = 19/25$

Hence, the probability that the problem is solved is 19/25.

11) A can hit a target 3 times in 5 shots, B hits target 2 times in 5 shots, C hits target 3 times in 4 shots. Find the probability of the target being hit when all of them try.

Solution) Let, P(A) denote the probability of A hitting the target.

P(B) denote the probability of B hitting the target.

P(C) denotes the probability of C hitting the target.

Given, $P(A) = 3/5$, $P(B) = 2/5$, $P(C) = 3/4$, then the complementary events are

$P(A^1) = 2/5$, $P(B^1) = 3/5$, $P(C^1) = 1/4$

The probability of A, B, C all not hitting the target is $P(A^1 \cap B^1 \cap C^1)$:

$P(A^1 \cap B^1 \cap C^1) = P(A^1) . P(B^1) . P(C^1)$

⇨ $P(A^1 \cap B^1 \cap C^1) = 2/5 \cdot 3/5 \cdot 1/4$

⇨ $P(A^1 \cap B^1 \cap C^1) = 3/50$

$P(A \cup B \cup C)$ = The probability of at least one of A, B, C hitting the target the target.

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

⇨ $P(A \cup B \cup C) = 1 - 3/50$

⇨ $P(A \cup B \cup C) = 47/50$

BAYE'S THEOREM (RULE OF INVERSE PROBABILITY):

Statement: If an event 'A' can only occur in conjunction with one of the 'n' mutually exclusive and exhaustive events $E_1, E_2 \ldots\ldots E_n$ and if 'A' actually happens, then the probability that it was preceded by the particular event E_i (i=1, 2……n) is given by:

$$P(E_i/A) = \frac{P(A \cap E_i)}{\Sigma P(E_i) \cdot P(A/E_i)} = \frac{P(E_i) \cdot P(A/E_i)}{\Sigma P(E_i) \cdot P(A/E_i)}$$

Proof:

Since, the event A can occur in conjunction with any one of the mutually exclusive and exhaustive events $E_1, E_2 \ldots\ldots E_n$ we have:

$A = (A \cap E_1) \cup (A \cap E_2) \cup \ldots\ldots \cup (A \cap E_n)$

Where, $A \cap E_1, A \cap E_2 \ldots\ldots\ldots A \cap E_n$ being the subsets of mutually exclusive events $E_1, E_2 \ldots\ldots E_n$ are all disjoint (mutually exclusive) events.

Hence, by the multiplication theorem of probability, we have:

$P(A) = P\{(A \cap E_1) \cup (A \cap E_2) \cup \ldots\ldots \cup (A \cap E_n)\}$

$\Rightarrow P(A) = P(A \cap E_1) + P(A \cap E_2) + \ldots\ldots\ldots + P(A \cap E_n)$

$\Rightarrow P(A) = P(E_1) \cdot P(A/E_1) + P(E_2) \cdot P(A/E_2) + \ldots\ldots\ldots + P(E_n) \cdot P(A/E_n)$

$\Rightarrow P(A) = P(E_i) \cdot P(A/E_i) \rightarrow (1)$

For any particular event E_i, the conditional probability, $P(E_i/A)$ is given by:

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

This is the Baye's rule for obtaining conditional probabilities.

Hence, Baye's theorem is proved.

PROBLEMS ON BAYE'S THEOREM:

1) In a certain college 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body. (a) What is the probability that mathematics is being studied? (b) If a student is selected at random and is found to be studying mathematics, find the probability that the student is a girl? (c) A boy?

Solution) Let, B = the event of selecting a boy.

G = the event of selecting a girl.

M = the event that a student is studying mathematics.

Given that, the probability of selecting a boy is P(B) = 40/100 = 2/5

The probability of selecting a girl is P(G) = 60/100 = 3/5

Probability that mathematics is studied, given the student is a boy:

P(M/B) = 25/100 = 1/4

Probability that mathematics is studied, given the student is a girl:

P(M/G) = 10/100 = 1/10

(a) Probability that maths is studied is P(M) = P(G) . P(M/G) + P(B) . P(M/B)

Therefore, by total probability theorem, we have,

P(M) = (3/5) . (1/10) + (2/5) . (1/4) = 4/25

(b) By Baye's theorem, probability of mathematics student is a girl is P(G/M).

$P(G/M) = \frac{P(G) \cdot P(M/G)}{P(M)} = \frac{(3/5)(1/10)}{(4/25)}$

Therefore, P(G/M) = 3/8.

(c) By Baye's theorem, probability of mathematics student is a boy is P(B/M).

$P(B/M) = \frac{P(B) \cdot (M/B)}{P(M)} = \frac{(2/5)(1/4)}{(4/25)}$

Therefore, P(B/M) = 5/8.

2) The chance that doctor A will diagnose a disease 'x' correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. The patient of doctor A, who had disease 'x', died. What is the chance that his disease was diagnosed correctly?

Solution) Let E_1 be the event that "disease 'x' is diagnosed correctly by doctor A" and E_2 be the event that "a patient of doctor A who has disease 'x' died".

Then, $P(E_1) = 60/100 = 0.6$, and $P(E_2 / E_1) = 40/100 = 0.4$

Also, $P(E_1^c) = 1 - P(E_1) = 1-0.6 = 0.4$, and $P(E_2/E_1^c) = 70/100 = 0.7$

Therefore, by Baye's theorem, we have:

$$P(E_1/E_2) = \frac{P(E_1) \cdot P(E_2/E_1)}{P(E_1)P(E_2/E_1) + P(E_1^c)P(E_2/E_1^c)}$$

$$\Rightarrow P(E_1/E_2) = \frac{0.6 \times 0.4}{0.6x0.4 + 0.4x0.7} = 6/13.$$

Hence, the probability of chance that the disease was diagnosed correctly is 6/13.

3) A bag 'A' contains 2 white and 3 red balls and bag 'B' contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that the red ball drawn is from bag 'B'?

Solution) Let E_1, E_2 be the two events of drawing a ball from bags 'A' and bag 'B' respectively.

The probability of selecting bag 'A' from the two bags is $P(E_1) = 1/2$.

The probability of selecting bag 'B' from the two bags is $P(E_2) = 1/2$.

Let, 'R' be the event of drawing a red ball from any of the bags.

Then, the probability of drawing a red ball from bag 'A' = $P(R/E_1) = 3/5$.

The probability of drawing a red ball from bag 'B' = $P(R/E_2) = 5/9$.

Therefore, the probability that a red ball is drawn from bag 'B' is $P(E_2/R)$.

i.e. $$P(E_2/R) = \frac{P(E_2) \cdot P(R/E_2)}{P(E_1) \cdot P(R/E_1) + P(E_2) \cdot P(R/E_2)}$$

⇨ $$P(E_2/R) = \frac{(1/2) \cdot (5/9)}{(1/2)(3/5) + (1/2) \cdot (5/9)}$$

⇨ $$P(E_2/R) = \frac{(5/9)}{(52/45)} = 25/52$$

Hence, the probability that the red ball drawn is from bag 'B' is 25/52.

4) First box contains 2 black, 3 red, 1 white ball; second box contains 1 black, 1 red, 2 white balls and third box contains 5 black, 3 red, 4 white balls. Of these a box is selected at random. From it a red ball is randomly drawn. If the ball is red, find the probability that is from second box.

Solution) Let x, y, z be first, second and third boxes.

Probability of selecting a box from three boxes is P(x) = 1/3.

Probability of selecting a box from three boxes is P(y) = 1/3.

Probability of selecting a box from three boxes is P(z) = 1/3.

Let R be the event of drawing a red ball from a box.

Then, P(R/x) = 3/6, P(R/y) = 1/4, P(R/z) = 3/12

Therefore, by Baye's theorem, the required probability is P(y/R):

$$P(y/R) = \frac{P(y) \cdot P(R/y)}{P(x) \cdot P(R/x) + P(y) \cdot P(R/y) + P(z) \cdot P(R/z)}$$

⇨ $P(y/R) = \frac{(1/3) \cdot (1/4)}{(1/3)(3/6) + (1/3)(1/4) + (1/3)(3/12)} = 1/4$

Hence, the probability of drawing a red ball form the second box is 1/4.

5) Suppose 5 men out of 100 and 25 women out of 10, 000 are color blind. A color blind person is chosen at random. What is the probability of the person being a male (Assume male and female to be in equal numbers).

Solution) Given that, 5 men out 100 and 25 women out of 10, 000 are color blind. A color blind person is chosen at random.

Then the probability that the chosen person is male is P(M) = 1/2 = 0.5.

Similarly, the probability that the chosen person is female is P(W) = 1/2 = 0.5.

Let B represent a blind person, then we have:

P(B/M) = 5/100 = 0.05 & P(W/M) = 25/10, 000 = 0.0025

The probability that the chosen person is male is P(M/B).

Therefore, $P(M/B) = \frac{P(M) \cdot P(B/M)}{P(M) \cdot P(B/M) + P(W) \cdot P(B/W)}$

⇨ $P(M/B) = \frac{(0.05) \times (0.5)}{(0.05)(0.5) + (0.0025)(0.5)}$

⇨ P(M/B) = 9.5

Hence, the probability of a color blind person being a male is 9.5

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

(i) Machine A. (ii) Machine B. (iii) Machine C.

Solution) Let P(A), P(B), P(C) denote the probabilities of the events "bolts manufactured by the machines A, B, C. Then by given data we have the following:

P(A) = 20/100 = 1/5, P(B) = 30/100 = 3/10, P(C) = 50/100 = 1/2

Let D denote that the bolt drawn is found to be defective, and then probabilities are:

P(D/A) = 6/100, P(D/B) = 3/100, P(D/C) = 2/100

(i) If bolt is defective, then the probability that it is from Machine A = P(A/D).

$P(A/D) = \frac{P(A) \cdot P(D/A)}{}$

$$P(A) \cdot P(D/A) + P(B) \cdot P(D/B) + P(C) \cdot P(D/C)$$

- $P(A/D) = \dfrac{(1/5) \cdot (6/100)}{(1/5) \cdot (6/100) + (3/10) \cdot (3/100) + (1/2) \cdot (2/100)}$

- $P(A/D) = 12/31.$

Hence, the probability that the selected defective bolt manufactured by Machine A is 12/31.

(ii)) If bolt is defective, then the probability that it is from Machine A = P(A/D).

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

- $P(B/D) = \dfrac{(3/10) \cdot (3/100)}{(1/5) \cdot (6/100) + (3/10) \cdot (3/100) + (1/2) \cdot (2/100)}$

- $P(B/D) = 9/31.$

Hence, the probability that the selected defective bolt manufactured by Machine B is 9/31.

(iii)) If bolt is defective, then the probability that it is from Machine A = P(A/D).

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

- $P(C/D) = \dfrac{(1/2) \cdot (2/100)}{(1/5) \cdot (6/100) + (3/10) \cdot (3/100) + (1/2) \cdot (2/100)}$

- $P(C/D) = 10/31.$

Hence, the probability that the selected defective bolt manufactured by Machine C is 10/31.

7) Of the three men, the chances that a politician, a businessman or an academician will be appointed as a vice-chancellor (V.C) of a university are 0.5, 0.3, and 0.2 respectively. Probabilities that research is promoted by these persons if they are appointed as V.C are 0.3. 0.7, 0.8 respectively.

(i) Determine the probability that research is promoted.

(ii) If research is promoted, what is the probability that V.C is an academician?

Solution) Let A, B, C are the three events that a politician, a businessman or an academician will be appointed as V.C of the three men. Their probabilities are given below:

$P(A) = 0.5 = 5/10, \quad P(B) = 0.3 = 3/10, \quad P(C) = 0.2 = 2/10$

Let R be the event that the research is promoted and appointed as V.C.

Then, the probabilities that research is promoted if they are appointed as V.Cs are:

$P(R/A) = 0.3 = 3/10,\ P(R/B) = 0.7 = 7/10,\ P(R/C) = 0.8 = 8/10$

(i) The probability that the research is promoted is = $P(R/A) + P(R/B) + P(R/C)$

- $(3/10) + (7/10) + (8/10)$
- $18/10 = 1.8$

Hence, the probability that research is promoted either by a politician or a businessman or an academician is 1.8

(ii) The probability that if research is promoted that the V.C is an academician is:

$$P(C/R) = \frac{P(C) \cdot P(R/C)}{P(A) \cdot P(R/C) + P(B) \cdot P(R/B) + P(C) \cdot P(R/C)}$$

- $$P(C/R) = \frac{(2/10) \cdot (8/10)}{(5/10) \cdot (3/10) + (3/10) \cdot (7/10) + (2/10) \cdot (8/10)}$$
- $P(C/R) = 4/13.$

Hence, the probability that if research is promoted that the V.C is an academician is 4/13.

8) A business man goes to hotels X, Y, Z, 20%, 50%, and 30% of the time respectively. It is known that 5%, 4%, 8% of the rooms in X, Y, Z hotels have faulty plumbing. What is the probability that business man's room having faulty plumbing is assigned to hotel Z.

Solution) Let the probabilities of business man going to hotels X, Y, Z are respectively P(X), P(Y), P(Z). i.e.,

$P(X) = 20/100 = 2/10, \quad P(Y) = 50/100 = 5/10,\ P(Z) = 30/100 = 3/10$

Let E = the event that the hotel room having faulty plumbing.

The probabilities that X, Y, Z hotels have faulty plumbings are:

$P(E/X) = 5/100 = 1/20,\ P(E/Y) = 4/100 = 1/25,\ P(E/Z) = 8/100 = 2/25$

The probability that the business man's room having faculty plumbing is assigned to hotel Z is P(Z/E).

$$P(Z/E) = \frac{P(Z) \cdot P(E/Z)}{P(X) \cdot P(E/X) + P(Y) \cdot P(E/Y) + P(Z) \cdot P(E/Z)}$$

$$\Rightarrow \quad P(Z/E) = \frac{(2/25) \cdot (3/10)}{(2/25) \cdot (3/10) + (1/25) \cdot (5/10) + (1/20) \cdot (2/10)}$$

$$\Rightarrow \quad P(Z/E) = 4/9$$

Hence, the probability that business man's room having faulty plumbing assigned to hotel Z is 4/9.

9) There are two boxes. In box I, 11 cards are there numbered 1 to 11 and in the box II, 5 cards are there numbered 1 to 5. A box is chosen and a card is drawn. If the card shows an even number then another card is drawn from the same box. If card shows an odd number another card is drawn from the other box. Find the probability that (i) both are even (ii) both are odd (iii) if both are even, what is the probability that they are from box I.

Solution)

Number of cards in box I = 11

Number of cards with even numbers in box I = 5

Number of cards with odd numbers in box I = 6

Number of cards in box II = 5

Number of cards with even numbers in box II = 2

Number of cards with odd numbers in box II = 3

The probability of choosing any one box is = 1/2

(i) Let E = the event that both the cards are even.

For this a box is chosen and a card is picked, if the first card is even then the second card is also picked from the same box and that card is also even.

Let E_1 = both the cards are from box I.

Therefore, $P(E_1) = (1/2) \cdot (5/11) \cdot (4/10) = 1/11$

Let E_2 = both the cards are from box II.

Therefore, $P(E_2) = (1/2) \cdot (2/5) \cdot (1/4) = 1/20$

Then, $P(E) = P(E_1) + P(E_2) = (1/11) + (1/20) = 31/220$

$$\Rightarrow \quad P(E) = 31/220$$

Hence, the probability that both the cards are even is 31/220.

(ii) Let E= the event that both the cards are odd.

For this a box is chosen and a card is picked, if the first card is odd then the second card is also picked from another box and that card is also odd.

Let E_1 = first card is from box I and second card is odd from box II.

Therefore, $P(E_1) = (1/2) . (6/11) . (3/5) = 9/55$

Let E_2 = first card is odd from box II and second card is odd from box I.

Therefore, $P(E_2) = (1/2) . (3/5) . (6/11) = 9/55$

Then, $P(E) = P(E_1) + P(E_2) = 9/55 + 9/55 = 18/55$

⇨ $P(E) = 18/55$

Hence, the probability that both the cards are odd is 18/55.

(iii) The probability that both cards are even and from box I is $P(E_1)$.

$P(E_1) = (1/2) . (5/11) . (4/10) = 1/11$

The probability that both cards are even and from box II is $P(E_2)$.

$P(E_2) = (1/2) . (2/5) . (1/4) = 1/20$

By using Baye's theorem, the probability that if both cards are even then they are from box I is:

⇨ $\frac{P(E_1)}{P(E_1) + P(E_2)}$

⇨ $\frac{(1/11)}{(1/2) . (5/11) . (4/10) + (1/2) . (2/5) . (1/4)}$

⇨ $\frac{(1/11)}{1/11 + 1/20}$

⇨ 20/31

Hence, the probability that both are even and are from box I is 20/31.

10) The bolts are drawn from a box containing 4 good and 6 bad balls. Find the probability that the second bolt is good if the first one is found to be bad.

Solution)

Let G = Event of getting a good bolt.

B = Event of getting a bad bolt.

Then, P(B) = Probability of getting bad bolt. i.e., P(B) = 6/10

P(G) = Probability of getting good bolt. i.e., P(G) = 4/9

P(G/B) = Probability of getting second bolt good given the first bolt is bad.

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

⇨ $P(G \cap B) = P(B) \cdot P(G/B)$

⇨ $P(G \cap B) = (6/10) \cdot (4/9) = 4/15$

⇨ $P(G \cap B) = 4/15$

Then, $P(G/B) = \frac{4/15}{6/10}$

⇨ $P(G/B) = 4/9.$

Hence, the probability that second bolt is good if the first bolt is found to be bad is 4/9.

11) In a factory, machine 'A' produces 40% of the output and machine 'B' produces 60%. On the average, 9 items in 1000 produced by 'A' are defective and 1 item in 250 produced by 'B' is defective. An item drawn at random from a day's output is defective. What is the probability that it was produced by A'' or 'B'?

Solution)

Output produced by machine A is = 40%, then the probability of machine A is

$$P(A) = 40/100 = 0.4$$

Output produced by machine B is = 60%, then the probability of machine B is

$$P(B) = 60/100 = 0.6$$

Let 'D' be the event of getting the item as defective item.

P(D/A) = Probability that items produced by 'A' are defective = 9/1000 = 0.009

Similarly, P(D/B) = probability of items produced by 'B' are defective = 1/250 = 0.004

(i) Then, P(A/D) = Probability of manufacturing the defective bolt by machine A.

$$\text{Therefore, } P(A/D) = \frac{P(A) \cdot P(D/A)}{P(A) \cdot P(D/A) + P(B) \cdot P(D/B)}$$

- $P(A/D) = \frac{0.4 \times 0.009}{0.4 \times 0.009 + 0.6 \times 0.004}$
- $P(A/D) = 0.0036/0.006 = 0.6$
- $P(A/D) = 0.6$

Hence, the probability of manufacturing the defective bolt by machine A is 0.6.

(ii) Then, P(B/D) = Probability of manufacturing the defective bolt by machine B.

Therefore, $P(B/D) = \frac{P(B) \cdot P(D/B)}{P(A) \cdot P(D/A) + P(B) \cdot P(D/B)}$

- $P(B/D) = \frac{0.6 \times 0.004}{0.4 \times 0.009 + 0.6 \times 0.004}$
- $P(B/D) = 0.0024/0.006 = 0.4$
- $P(B/D) = 0.4$

Hence, the probability of manufacturing the defective bolt by machine B is 0.4.

UNIT-II

PROBABILITY DISTRIBUTIONS

THEORITICAL DISTRIBUTIONS:

INTRODUCTION:

The Statistical measures like the averages, dispersion, skew ness, kurtosis, correlation etc., for the sample frequency distributions, not only give us the nature and form of the sample data but also help us in formulating certain ideas about the population characteristics. However a more scientific way of drawing inferences about the population characteristics is through the study of theoretical distribution which we shall discuss in this chapter.

In the population the values of the variable may be distributed according to some definite probability law which can be expressed mathematically and the corresponding probability distribution is known as **Theoretical Probability Distribution.**

In this chapter we shall study the following theoretical distributions:

i. Binomial Distribution

ii. Poisson Distribution

iii. Normal Distribution

The first two distributions are discrete probability distributions and the third is a continuous probability distribution.

BINOMIAL DISTRIBUTION:

Binomial distribution is also known as the **Bernouli Distribution**, after the Swiss Mathematician **James Bernouli**, who discovered the Binomial distribution. This distribution can be used under the following conditions:

1. 'n' is the number of trials in finite and fixed.
2. The outcome of each trial may be classified into two mutually disjoint categories called "success and failure".
3. The probability of success in any trial is 'p' and is constant for each trial, $q = 1 - p$ i.e. $p + q = 1$, is then termed as probability of a success and a failure is constant for each trial.
4. All the trials are independent.

Note: The trials satisfying the above four conditions are also known as Bernouli trials.

PROBABILITY FUNCTION OF BINOMIAL DISTRIBUTION:

If 'X' denotes the number of successes in 'n' trials satisfying the characteristics of binomial distribution, then 'X' is a random variable which can take the values 0, 1, 2 …. N, then the general expression for the probability of 'X' success is given by:

$$P(x) = P(X = x) = {}^{n}c_{x}\, p^{x}\, q^{n-x};\ x = 0, 1, 2 \ldots n$$

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

Where, n = Total number of trials

x = Number of Successes, n – x = Number of Failures

p = Probability of Success, q = Probability of Failure

Properties:

1. The mean of the binomial distribution is "np".
2. Variance of the binomial distribution is "npq".
3. Standard Deviation of the binomial distribution is $\sqrt{npq}$.

Note: Binomial distribution is denoted by $X \sim B(n, p)$ ➔ 'X' follows binomial distribution with parameters n, p.

FITTING OF BINOMIAL DISTRIBUTION:

Suppose a random experiment consists of 'n' trials, satisfying the conditions of binomial distribution and suppose if this experiment is repeated 'N' times then the frequency of 'n' success is given by the formula:

$N * P(x) = N * nc_x p^x q^{n-x}$, where x = 0, 1, 2 … n

By putting x = 0, 1, 2 … n we get the expected or theoretical frequencies of binomial distribution. We first find the mean of the given frequency distribution by the formula:

$x = \Sigma (fx / N)$, is equal to np, which is the mean of the binomial distribution.

Therefore, np = x ➔ p = x/n, then q = 1 – p.

With these values of 'p' and 'q' the expected or theoretical frequencies of binomial distribution can be obtained.

POISSON DISTRIBUTION:

Poisson distribution was derived by a French Mathematician the Poisson. Poisson distribution may be obtained as a limiting case of binomial probability distribution under the following conditions:

1. 'n', the number of trials is indefinitely large i.e. n → ∞.
2. 'p', the constant probability of success for each trial is indefinitely small i.e. p → 0.
3. np = m (say) is finite.

DEFINITION:

The probability mass function of Poisson distribution $p(x) = \frac{e^{-m} m^x}{x!}$; x = 0, 1, 2 … n

Where, x = Number of successes, m = Average E = A mathematical constant

Properties:

1. Parameter of Poisson distribution is 'm'.
2. Mean of the Poisson distribution is 'm'.
3. Variance of the Poisson distribution is 'm'.
4. Standard deviation of the Poisson distribution is √m.

Note: The Poisson distribution is denoted by X ~ p (m) (X follows Poisson distribution with parameter 'm').

IMPORTANCE OF POISSON DISTRIBUTION:

The Poisson distribution can be used to explain the behavior of the discrete random variables where the probability of occurrence of the event is very small and the total number of possible cases is sufficiently large. The following are the some practical situations where Poisson distribution can be used.

1. The number of wrong telephone calls arriving at a telephone switch board in unit time (say one minute).

2. The number of customers arriving at the super market. (Say per hour).

3. The number of defective material in a packing manufactured by a good concern.

4. The number of suicides reported in a particular day or the number of persons dying due to a rare diseases such as heart attack, cancer or snake bite in a year.

FITTING OF POISSON DISTRIBUTION:

If we want to fit a Poisson distribution to a given frequency distribution. We compute the mean x of the given distribution and take it equal to the mean of the Poisson distribution i.e. we take m = x. Once 'm' is known as the various probabilities of the Poisson distribution we can obtain the probability mass function of the Poisson distribution as follows:

$$p(x) = \frac{e^{-m} m^x}{x!} ; x = 0, 1, 2 \ldots n$$

The theoretical or expected frequencies are obtained by using the following formula:

$f(x) = N\, p(x)$, where, N = Total observed frequencies.

NORMAL DISTRIBUTION:

Normal probability distribution commonly called the Normal distribution is one of the most important continuous theoretical distributions in statistics. Most of the data relating to economic and business statistics or even in social and physical sciences confirm through this distribution.

DEFINITION:

If 'x' is a continuous random variable, following normal probability distribution with mean 'μ' and standard deviation 'σ', then its probability density function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} [e^{(-1/2)[(x-\mu)/\sigma]^2}, \quad \text{where, } -\infty < x < \infty$$

Where, 'Π' & 'e' are mathematical constants, μ = Mean of Normal distribution,

σ = Standard deviation, x = Continuous random variable

Note: Parameters of normal distribution are mean 'μ' and standard deviation 'σ'.

STANDARD NORMAL VARIATE:

If 'X' is a random variable following normal distribution with mean 'μ' and standard deviation 'σ', then the Standard Normal Variate 'Z' is defined as follows:

$$Z = \frac{X - E(X)}{\sqrt{Var(x)}} = \frac{X - E(X)}{S.D.(x)} = \frac{X - \mu}{\sigma}$$

$Z = \frac{X - \mu}{\sigma}$, is called Standard Normal Variate.

Note: Parameters of Standard Normal Variate are mean '0' and standard deviation '1'.

RELATION BETWEEN BINOMIAL & NORMAL DISTRIBUTIONS:

Normal distribution is a limiting case of the binomial probability distribution under the following conditions:

1. 'n', the number of trials is indefinitely large i.e. $n \rightarrow \infty$.
2. Neither 'p' nor 'q' is very small.

We know that for a binomial Variate 'x' with parameters 'n' and 'p'; E (x) = np (mean) and V (x) = npq (variance).

Under the above two conditions the distributions of standard binomial Variate is:

$$Z = \frac{X - E(X)}{\sqrt{Var(x)}} \rightarrow Z = \frac{X - np}{\sqrt{npq}}$$

RELATION BETWEEN POISSON & NORMAL DISTRIBUTIONS:

If 'x' is a random variable following Poisson distribution with parameter 'm' then E (x) = m (mean) and V (x) = m (variance). Thus the Standard Normal Variate becomes as:

$$Z = \frac{X - E(X)}{\sqrt{Var(x)}} \rightarrow Z = \frac{X - m}{\sqrt{m}}$$

Normal distribution may also be regarded as a limiting case of Poisson distribution as the parameter $m \rightarrow \infty$.

PROPERTIES OF NORMAL DISTRIBUTION:

1. The shape of the Normal distribution is bell shaped curve.
2. The curve is symmetrical about the line $x = \mu$ i.e. it has the same shape on the either side of the line $x = \mu$.
3. Since the distribution is symmetrical, mean, median and mode are equal. Thus Mean = Median = Mode = μ.
4. The total area under the normal curve is one; the area to the right hand side as well as the left hand side is 0.5.
5. For normal distribution, quartiles are equidistant from second quartile (median).
6. No portion of the curve lies below the X – axis, since f (x) being the probability can ever be negative.
7. As 'x' increases numerically the value of f (x) decreases rapidly.
8. Normal distribution is a uni-model the only made occurring at $x = \mu$.

AREA PROPERTY:

One of the most fundamental properties of the normal probability curve is the Area Property. The area under the normal probability curves between the point's $(\mu - \sigma)$ to $(\mu + \sigma)$ is 68.26% or 0.6826.

The area under the normal probability curve between the points $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$ is 95.44% or 0.9544.

The area under the normal probability curve between the points $(\mu - 3\sigma)$ to $(\mu + 3\sigma)$ is 99.73% or 0.9973.

UNIT – III: SAMPLING THEORY

SAMPLE DESIGN AND SAMPLING PROCEDURES

SAMPLE DESIGN:

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design may as well lay down the number of times to be included in the sample i.e., the size of the sample. Sample design is determined before data are collected. There are many sample designs from which a researcher can choose. Some designs are relatively more precise and easier to apply than others. Researcher must select/prepare a sample design which should be reliable and appropriate for his research study.

STEPS IN SAMPLE DESIGN:

While developing a sample design, the researcher must pay attention to the following points:

1. ***Type of universe:*** The first step in developing sample design is to clearly define the set of objects, technically called the Universe, to be studied. The universe can be finite or infinite. In finite universe the number of items is certain, but in case of an infinite universe the number of items is infinite i.e., we cannot have any idea about the total number of items. The population of a city, the number of workers in a factory and the like are examples of finite universes, whereas the number of stars in the sky, listeners of a specific radio programme, throwing of a dice etc., are examples of infinite universes.
2. ***Sampling Unit:*** A decision has to be taken concerning a sampling unit before selecting sample. Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual. The researcher will have to decide one or more of such units that he has to select for his study.
3. ***Source List:*** It is also known as „Sampling frame" from which sample is to be drawn. It contains the names of all items of a universe (in case of finite universe only). If source list is not available, researcher has to prepare it. Such a list should be comprehensive, correct, reliable and appropriate. It is extremely important for the source list to be as representative of the population as possible.
4. ***Size of sample:*** This refers to the number of items to be selected from the universe to constitute a sample. This major problem before a researcher. The size of sample should neither be excessively large, nor too small. It should be optimum. An optimum sample is one which fulfills the requirements of efficiency, representative-ness, reliability and flexibility. While deciding the size of sample, researcher must determine the desired precision as also an acceptable confidence level for the estimate.
5. ***Parameters of interest:*** In determining the sample design, one must consider the question of the specific population parameters which are of interest. For instance, we may be interested in estimating the proportion of persons with some characteristic in the population, or we may be interested in knowing some average or the other measure concerning the population. They are also important sub-groups in the population about whom we would like to make estimates. All this has a strong impact upon the sample design we would accept.
6. ***Budgetary Constraint:*** Cost considerations, from practical point of view, have a major impact upon decisions relating to not only the size of the sample but also to the type of sample. This fact can even lead to the use of a non-probability sample.
7. ***Sampling Procedure:*** Finally, the researcher must decide the type of sample he will use i.e., he must decide about the technique to be used in selecting the items for the sample. In fact, this technique or procedure stands for the sample design itself. There are several sample designs out of which the researcher must choose one for his study. Obviously, he must select that design which, for a given sample size and for a cost, has a small sampling error.

CHARACTERISTICS OF GOOD SAMPLE DESIGN:

From what has been stated above, we can list down the characteristics of a good sample design as under:

a) Sample design must result in a truly representative sample.

b) Sample design must be such which results in a small sampling error.

c) Sample design must be viable in the context of funds available for the research study.

d) Sample design must be such so that systematic bias can be controlled in a better way.

e) Sample should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence.

CRITERIA OF SELECTING A SAMPLING PROCEDURE:

In this context one must remember that two costs are involved in a sampling analysis viz., the cost of collecting the data and the cost of an incorrect inference resulting from the data. Researcher must keep in view the two causes of incorrect inferences viz., systematic bias and sampling error. Systematic bias results from errors in the sampling procedures, and it cannot be reduced or eliminated by increasing the sample size. At best the causes responsible for these errors can be detected and corrected. Usually a systematic bias is the result of one or more of the following factors.

1) Inappropriate frame: If the sampling frame is inappropriate i.e., a biased representation of the universe, it will result in a systematic bias.

2) Defective measuring device: If the measuring device is constantly in error, it will return in systematic bias. In survey work, systematic bias can result if the questionnaire or the interviewer is biased. Similarly, if the physical measuring device is defective there will be systematic bias in the data collected through such a measuring device.

3) Non-respondents: If we are unable to sample all the individuals initially include in the sample, there may arise a systematic bias. The reason is that in such a situation the likelihood of establishing contact or receiving a response from an individual is often correlated with the measure of what is to be estimated.

4) Indeterminacy principle: Sometimes we find that individuals act different when kept under observation that what they do when kept in non-observed situations. For instance, if workers are aware that somebody is observing then in course of a work study on the basis of which the average length of time to complete a task will be determined and accordingly the quota will be set for piece work, they generally tend to work slowly in comparison to the speed with which they work if kept unobserved. Thus, the indeterminacy principle may also be a cause of a systematic bias.

5) **Natural bias in the reporting of data:** Natural bias of respondents in the reporting of data is often the cause of a systematic bias in many inquiries. There is usually a download bias in the income data collected data by government taxation department, whereas we find an upward bias in the income data collected by some social organization. People in general understate their incomes if asked about it for tax purposes, but they overstate the same if asked for social status or their affluence. Generally in psychological surveys, people tend to give what they think is the „correct" answer rather than revealing their true feelings.

DIFFERENT TYPES OF SAMPLE DESIGNS:

There are different types of sample designs based on two factors viz., the representation basis and the element selection technique. On the representation basis and the element selection technique. On the representation basis, the sample may be probability sampling or it may be non-probability sampling. Probability sampling is based on the concept of random selection, whereas non-probability sampling is „non-random sampling. On element selection bias, the sample may be either unrestricted or restricted. When each sample element is drawn individually from the population at large, then the sample so drawn is known as „unrestricted sample", whereas all other forms of sampling are covered under the term „restricted sampling". The following chart exhibits the sample designs as explained above.

Non-probability sampling: Non-probability sampling is that sampling procedure which does not afford any basis for estimating the probability that each item in the population has of being included in the sample. Non-probability sampling is also known by different names such as deliberate sampling, purposive sampling and judgment sampling. In this type if sampling, items for the sample are selected deliberately by the researcher; his choice concerning the items remains supreme. In other words, under non-probability sampling the organizers of the inquiry purposively choose the particular units of the universe for consulting a sample on the basis that the small mass that they so select out of a huge one will be typical or representative of the whole. For instance, if economic conditions of people living in a state are to be studied, a few towns and villages may be purposively selected for intensive study on the principle that they can be representative of the entire state. Thus, the judgment of the organizers of the study plays an important part in this sampling design.

Quota sampling: It is also an example of non-probability sampling. Under quota sampling the interviewers are simply given quotas to be filled from the different strata, with some restrictions on how they are to be filled. In other words, the actual selection of the items for the sample is left to the interviewer"s discretion. This type of sampling is very convenient and is relatively inexpensive. But the samples so selected certainly do not possess the characteristic of random samples. Quota samples are essentially judgment samples and inferences drawn on their basis are not amenable to statistical treatment in a formal way.

Probability sampling: Probability sampling is also known as „random sampling" or „chance sampling". Under this sampling design, every time of the universe has an equal chance of inclusion in the sample. It is, so to say, a lottery method in which individual units are picked up from the whole group not deliberately but by some mechanical process. Here it is blind chance alone that determines whether one item or the other is selected. The results obtained from probability or random sampling can be assured in terms of probability i.e., we can measure the errors of estimation or the significance of results obtained from a random sample, and this fact brings out the superiority of random sampling design over the deliberate sampling design.

Random sampling ensures the Law of Statistical Regularity which states that if on an average the sample chosen is a random one, the sample will have the same composition and characteristics as the universe. This is the reason why random sampling is considered as the best technique of selecting a representative sample.

Random sampling from a finite population to that method of sample selection which gives each possible sample combination an equal probability of being picked up and each item in the entire population to have an equal chance of being included in the sample. This applies to sampling without replacement i.e., once an selected for the sample, it cannot appear in the sample again (sampling with replacement is used less frequently in which procedure the element for the sample is returned to the population before the next element is selected. In such a situation the same element could appear twice in the same sample before the second element is chosen).in brief, the implications of random sampling (or simple random sampling) are:

(a) It gives each element in the population an equal probability of getting into the sample; and all choices are independent of one another.

(b) It gives each possible sample combination an equal probability of being chosen.

COMPLEX RANDOM SAMPLING DESIGNS:

Probability sampling under restricted sampling techniques, as stated above, may result in complex random sampling designs. Such designs may as well be called „mixed sampling designs" for many of such designs may represent a combination of probability and non-probability sampling procedures in selecting a sample. Some of the popular complex random sampling designs are as follows:

(i) Systematic Sampling: In some instances, the most practical way of sampling is to select every i^{th} item on a list. Sampling of this type is known as systematic sampling. An element of randomness is introduced into this kind of sampling by using random numbers to pick up the unit with which to start. For instance,

if a 4 percent sample is desired, the first item would be selected randomly from the first twenty-five and thereafter every 25^{th} item would automatically be included in the sample. Thus, in systematic sampling only the first unit is selected randomly and the remaining units of the sample are selected at fixed intervals. Although a systematic sample is not a random sample in the strict sense of the term, but it is often considered reasonable to treat systematic sample as if it were a random sample.

(ii) Stratified Sampling: If a population from which a sample is to be drawn does not constitute a homogeneous group, stratified sampling technique is generally applied in order to obtain a representative sample. Under stratified sampling the population is divided into several sub-populations that are individually more homogeneous than the total population a (the different sub-populations are called „strata") and then we select items from each stratum to constitute a sample. Since each stratum is more homogeneous than the total population, we are able to get precise estimates for each stratum and by estimating more accurately each of the component parts; we get a better estimate of the whole. In brief, stratified sampling results in more reliable and detailed information.

(iii) Cluster Sampling: If the total area of interest happens to be a big one , a convenient way in which a sample can be taken is to divide the area into a number of smaller non-overlapping areas and then to randomly select a number of these smaller areas (usually called clusters), with the ultimate sample consisting of all (or samples of) units in these small areas of clusters.

Thus in cluster sampling the total population is divided into a number of relatively small subdivisions which are themselves clusters of still smaller units and then some of these clusters are randomly selected for inclusion in the overall sample. Suppose we want to estimate the proportion of machine parts in an inventory which are defective. Also assume that there are 20000 machine parts in the inventory at a given point of time, stored in 400 cases of 50 each. Now using a cluster sampling, we would consider the 400 cases as clusters and randomly select „n" cases and examine all the machine parts in each randomly selected case.

Cluster sampling, no doubt, reduces cost by concentrating surveys in selected surveys. But certainly it is less precise than random sampling. There is also not as much information in „n" observations within a cluster as there happens to be in „n" randomly drawn observations. Cluster sampling is used only because of the economic advantage it possesses; estimates based on cluster samples are usually more reliable per unit cost.

(iv) Area Sampling: If clusters happen to be some geographic subdivisions, in that case cluster sampling is better known as area sampling. In other words, cluster designs, where the primary sampling unit represents a cluster of units based on geographic area, are distinguished as area sampling. The plus and minus points of cluster sampling are also applicable to area sampling.

(v) Multi-stage Sampling: Multi-stage sampling is a further development of the principle of cluster sampling. Suppose we want to investigate the working efficiency of nationalized banks in India and we want to take a sample of few banks for this purpose. The first stage is to select large primary sampling unit such as states in a country. Then we may select certain districts and interview all banks in the chosen districts. This would represent a two-stage sampling design with the ultimate sampling units being clusters of districts.

If instead of taking a census of all banks within the selected districts, we select certain towns and interview all banks in the chosen towns. This would represent a three-stage sampling design. If instead of taking a census of all banks within the selected towns, we randomly sample banks from each selected town, then it is a case of using a four-stage sampling plan. If we select randomly at all stages, we will have what is known as „multi-stage random sampling design".

Ordinarily multi-stage sampling is applied in inquires extending to a considerable large geographical area, say, the entire country. There are two advantages of this sampling design viz., (a) It is easier to administer than most single stage designs mainly because of the fact that sampling frame under multi-stage sampling in developed impartial units. (b) A large number of units can be sampled for a given cost under multistage because of sequential clustering, whereas this is not possible in most of the sample designs.

(vi) Sampling with probability proportional to size: In case the cluster sampling units do not have the same number or approximately the same number of elements, it is considered appropriate to use a

random selection process where the probability of each cluster being included in the sample is proportional to the size of the cluster. For this purpose, we have to list the number of the elements in each cluster irrespective of the method of ordering the cluster. Then we must sample systematically the appropriate number of elements from the cumulative totals.

(vii) Sequential Sampling: This sampling design is some what complex sample design. The ultimate size of the sample under this technique is not fixed in advance, but we determined according to mathematical decision rules on the basis of information yielded as survey progresses. This is usually adopted in case of acceptance sampling plan in context of statistical quality control. When a particular lot is to be accepted or rejected on the basis of single sample, it is known as single sampling; when the decision is to be taken on the basis of two samples, it is known as double sampling and in case the decision rests on the basis of more than two samples but the number of samples in certain and decide in advance, the sampling is known as the multiple sampling. But when the number of samples is more than two but it is neither certain nor decides in advance, this type of system is often referred to as sequential sampling.

Theory of Estimation:

Introduction:

In this theory of estimation we shall develop the technique which enables us to generalize the results of the sample to the population to find how far these generalizations are valid and also to estimate the population parameters along with degree of confidence. The answers to these and many other related problems are provided by a very important branch of statistics is known as Statistical Inference, which may be broadly classified in to the following two heads.

1. Theory of Estimation.
2. Testing of Hypothesis.
3. Estimation of population parameters like mean, variance etc. from the corresponding sample statistics is one of the very important problems of statistical inference.
4. The estimation of population parameter are imperative in making business decisions.
5. The theory of estimation was founded by Professor. R.A. Fisher in a series of fundamental papers round about 1930 and it is divided into two groups.
6. Point estimation.
7. Interval estimation.
8. In point estimation a sample statistics is used to provide an estimate of the population parameter whereas in interval estimation, probable range is specified within which the true value of parameter might be expected to lie.

TESTING OF HYPOTHESIS:

Introduction: The modern theory of probability plays a very important role and the branch of statistics which helps us in arriving at the criteria for such decisions is known as testing of hypothesis.

Tests of Significance: Accordingly a procedure to access the significance of a statistics (or) difference between two independent statistics is known as a test of significance.

Null Hypothesis: Null hypothesis is which is tested for possible rejection under the assumption that it is true. It is usually denoted by H_o.

Alternative Hypothesis: Any hypothesis which is completely to the null hypothesis is called an alternative hypothesis. It is usually denoted by H_1 (or) H_a.

Hypothesis:

- Any valuable statement about a population is called hypothesis.

Eg: The pass % of a college is sixty.

- Types of errors in testing of hypothesis.
- In testing of hypothesis inference consists in arriving at a decision to accept (or) reject a null hypothesis only a sample from it.
- In any test procedure, the four possible mutually disjoint and exhaustive decisions are

1. The H_0 is true and our test accepts it. (Correct decision).
2. The H_0 is false and our test rejects it (correct decision).
3. The H_0 is true and our test rejects it. (Wrong decision). Types of errors
4. The H_0 is false and our test of accepts it (wrong decision).

- In testing of hypothesis, we are likely to commit two types of errors.

1. Type-I error.
2. Type-II error.

Type-I error: The error of rejection H_0 is true is known as type 1 error. The probability of type 1 error is denoted by α. P [reject H_0 when it is true] = p [type 1 error] = α.

Type-II error: The error of accepting H_0 when H_0 is false is known as type 2 errors. The probability of type 2 error is denoted by β. P [accepting H_0 when it is false] = p [type 2 error] = β

- Then α and β are also called sizes of type 1 error and type 2 error respectively.
- In the terminology of industrial quality control while inspecting the quality of manufacturing lot (sample), the type 1 error amounts to rejecting a good lot and type 2 error amounts to accepting a bad lot.
- Accordingly: α = p [rejecting a good lot] β = p [accepting a bad lot]
- The sizes of type 1 and type 2 error are also known as producers risk and consumers risk respectively.

Level of significance: The maximum size of the type 1 error, which we are prepared to risk, is known as level of significance. It is denoted by α, and is given by p [rejecting H_0 when it is true]. Commonly used level of significance in practice is 5% (0.05) and 1% (0.01).

One-tail test and two-tail test:

One-tail test: Any statistical hypothesis is expressed with the symbol less than (<) or greater than (>), is called one-tail test. In one-tail test the entire rejection region lays only one side i.e., Right hand side (or) Left hand side.

Left-tail test: Any statistical hypothesis expressed with the symbol only "<" then it is called Left-hand test. In this left-tail test the entire critical region lien on left hand side.

Right-tail test: Any statistical hypothesis is expressed with the symbol only ">", then it is called right hand side. In this right hand side, the entire critical (rejection) region lies on the right hand side.

Two-tail test: Any statistical hypothesis is expressed with the symbol "≠", then the test is called two-tailed test. In this two-tailed test the entire critical region lien on both sides.

PROCEDURE FOR TESTING OF HYPOTHESIS:

- Set up the Null hypothesis (H_0).
- Set up the Alternative hypothesis (H_1).
- Alternative hypothesis will enable us to decide whether we have to use a one tail (right or left tail) (or) two-tailed test.

Level of significance: Choose the appropriate level of significance α.

Test Statistics:

- Define and compute the test statistics under H_0.
- Sum of the commonly used distributions in obtaining test statistics are normal Z, t, $\chi^{2,}$ F test.
- Obtain the table value (or) critical value of test statistics from the appropriate table.
- Compare these calculated values with the table values.
- If the calculated value of the test statistics lies in the rejection region, we reject H_0.
- If the computed value of test statistics lies outside the rejection region (acceptance region), we accept H_0.

E.g. $Z_{cal} \leq Z_{tab}$ → Accept H_0 $Z_{cal} > Z_{tab}$ → Reject H_0

LARGE SAMPLE TESTS:

- If sample size n>30, then the sample is called Large Samples.
- For large sample, there are four important tests for testing significant level.
- These are the tests of significance for:

1. Single mean
2. Difference of two means
3. Single Proportion
4. Difference of two Proportions

TEST OF SIGNIFICANCE FOR SINGLE MEAN:

If we want to test, whether the given sample of size 'n' has been drawn from a population mean 'μ', we said a null hypothesis i.e., there is no difference between sample mean X and population mean μ. Therefore, the test statistics corresponding to X is:

$$Z = \frac{X - \mu}{(\sigma/\sqrt{n})} \sim N(0, 1)$$

Where, X → sample mean μ → Population mean

σ → population S.D n → sample size.

After calculating the calculated value it compares with the tabulated value if it is ≤ accepted H_0 otherwise reject H_0.

TEST OF SIGNIFICANCE FOR DIFFERENCE OF TWO MEANS:

Let X is the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let X_2 be the mean of an independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2 .

Here the problem is:

1. To test the equality of the two population means i.e., $\mu_1 = \mu_2$ (or)
2. To test the significance of the difference between the independent samples mean X_1 X_2.

Basically both these problems are same. The corresponding null hypothesis will be set up as follows:

1. H_0: $\mu_1 = \mu_2$.
2. H_0: There is no significance difference between the sample means.

Under H_0, the test statistics is:

$$Z = \frac{X_1 - X_2}{} \sim N(0, 1)$$

$$\frac{\sigma_1^2}{\sqrt{n_1}} + \frac{\sigma_2^2}{\sqrt{n_2}}$$

Where, X_1 → first sample mean, X_2 → second sample mean,

σ_1^2 →first sample variance, σ_2^2 → second sample variance, n_1→first sample size, n_2→ second sample size.

In case of large samples,

$$Z = \frac{X_1 - X_2}{\frac{s_1^2}{\sqrt{n_1}} + \frac{s_2^2}{\sqrt{n_2}}} \sim N(0, 1)$$

After calculating the calculated value it compares with the table value if it is ≤accept H0, otherwise reject H0.

NOTE: If the samples are taken from the same population, with a common standard deviation i.e., σ12= σ22= σ then the test statistics becomes:

$$Z = \frac{X_1 - X_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

TEST OF SIGNIFICANCE FOR DIFFERENCE OF STANDARD DEVIATION:

If S_1 and S_2 are the S.D's of two independent samples then under the null hypothesis H_0: $\sigma_1 = \sigma_2$ i.e., both the population S.D's are equal.

The test statistics Z is:

$$Z = \frac{X_1 - X_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} \sim N(0, 1)$$

In case of large samples $\sigma_1^2 = S_1^2$ and $\sigma_2^2 = S_2^2$, then

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}} \sim N(0, 1)$$

Where, S_1 → first sample S.D. S_2 → second sample S.D.

n_1 →first sample size. n_2 →second sample size.

After calculating the calculated value, it compares with the table value, if it is less than (or) equal to accept H_0, otherwise reject H_0.

UNIT – IV: SMALL SAMPLE TEST, ANOVA & CHISQUARE

SMALL SAMPLE TESTS:

- If the size of the sample chosen from a population is ≤ 30, then the sample is called small sample.
- For small samples, there are four important tests, available for testing significance levels.
- These are the tests of significance for

1. Single mean.
2. Difference of significant mean (independent sample).
3. Difference of mean (dependent sample).
4. An observed correlation coefficient

DEGREES OF FREDOM:

- As the name suggest, the degrees of freedom abbreviated as degrees of freedom denotes the extent of independence (freedom) enjoyed by a given set of observation.
- Degrees of freedom are usually denoted by a Greek letter υ. **d.f (υ) = n-k.**

1. TESTS OF SIGNIFICANCE FOR SINGLE MEAN:

To test the null hypothesis, whether the sample mean and population differ significantly or not, then the test statistics is:

$t = \frac{X-\mu}{} \sim t_{n-1}$ d.f at α% L.O.S $S/\sqrt{n}$

Where, X□ sample mean. μ □ Population mean. S □ sample S.D. n □sample size n-1□degrees of freedom.

After calculating the calculated values it compares with the table if $t_{cal}<t_{tab}$ accept H_0 and $t_{cal}>t_{tab}$ reject H_0.

2. *TESTS OF SIGNIFICANCE FOR DIFFERENT BETWEEN TWO MEANS (DEPENDANT OBSERVATIONS/ PAIRED OBSERVATION):*

If the sample sizes are equal i.e., $n_1=n_2=n$. The two samples are dependant (or) related them. To test whether the samples mean differ significantly or not, the test statistics is:

$t = \frac{d}{}$ $\sim t_{n-1}$ d.f at α% L.O.S $S/\sqrt{n}$

Where, d □ $1/n \Sigma d_i$, d_i □ $X_i - Y_i$ S □ $\sqrt{(1/n-1) \Sigma (d_i-d)^2}$

After calculating calculated value, it compares with table value, if it is ≤ accept H_0 otherwise reject H_0.

3. TESTS OF SIGNIFICANCE FOR DIFFERENCE BETWEEN TWO MEANS (INDEPENDENTOBSERVATIONS):

Suppose we want to test two independent samples which have been drawn from two normal populations having the same means, and the population variances being equal. Let the Sample sizes be n_1, n_2.

Under the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, i.e. population variances are equal but unknown, then the test statistic will be:

$$t = \frac{x_1 - x_2}{S\sqrt{(1/n_1) + (1/n_2)}}$$

ANALYSIS OF VARIANCE

INTRODUCTION:

The **Analysis Of Variance** is a powerful statistical tool for tests of significance. The test of significance based on "t-distribution" is an adequate procedure only for testing the significance of the difference between two sample means. In a situation, when we have three or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the populations with the

same mean. **For Example:** Suppose five fertilizers are applied at random to four plots of the same shape and size and the yield of wheat on each of these plots is given. We may be interested in finding out whether the effect of these fertilizers on the yields is significantly different. The answer to this problem is provided by the technique of **Analysis of Variance**.

The basic purpose of the Analysis of Variance is to test the homogeneity of several means. The term analysis of variance was introduced by Professor R.A. Fisher in 1920 to deal with problems in analysis of Agronomical data. Variation (difference) is inherent in nature. The total difference in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes (ii) Chance (or) Random causes

The variation due to assignable causes can be identified and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be traced separately.

DEFINITION:

According to Professor R.A. Fisher, Analysis of Variance is the "separation of variance ascribable to one group of causes from the variance ascribable to the other group." By this technique the total variation in the sample date is expressed as the sum of its non-negative components.

ASSUMPTIONS OF ANOVA (Analysis Of Variance):

ANOVA test is based on the test statistic „F" (variance ratio test). For the validity of F-test in ANOVA the following assumptions are made:

- The observations are independent.
- Parent populations from which observations are taken are normal.
- Various treatments and environmental effects are additive in nature.
- There are two types of classifications in ANOVA.

1. ANOVA One-Way Classification 2. ANOVA Two-Way Classification

PROCEDURE OF ANOVA TEST:

Step-1: Set up the suitable Null Hypothesis. ($\mu_1 = \mu_2 = \mu_3$).
Step-2: Ste up the Alternative Hypothesis ($\mu_1 \neq \mu_2 \neq \mu_3$).
Step-3: Select the Test Statistic to be used. Here we use the test statistic as „F" distribution to test the hypothesis.
Step-4: Determine rejection and non-rejection regions.
Step-5: Calculate the value of the test statistic by using given data and compare with the table values. ***Step-6:*** After calculating the calculated value, compare that value with the table value. If the calculated value if "Less Than or Equal to (≤)" table value then accept the H_0 (Null Hypothesis) otherwise reject the H_0 (Null Hypothesis). $F_{cal} \leq F_{tab}$ (Accept H_0) $F_{cal} \geq F_{tab}$ (Reject H_0)

ANOVA ONE-WAY CLASSIFICATION:

The simplest form of Analysis of Variance is "**One-Way Classification**", which we use with simple random samples in order to compare the effect of single independent variable on the dependent variable. The scheme of classification according to single criteria is called One-Way classification and its analysis is known as "**One-Way Analysis of Variance**". The main objective of analysis of variance technique is to examine if there is a significant difference between the class Means in view of the inherent variability within the separate classes.

NULL HYPOTHESIS (H_0): We want to test the equality of population means i.e. the homogeneity of different groups. Therefore: **Null Hypothesis (H_0):** There is no significant difference between means of

„k" groups.
Therefore, $\mu_1 = \mu_2 = \ldots\ldots\ldots\ldots\ldots \mu_k = \mu$
ALTERNATIVE HYPOTHESIS (H_1): There is no significant difference between means of „k" groups. **STATISTICAL ANALYSIS: Total Variation = Variation due to Rows + Variation due to Error** Therefore, **Total Sum of Squares = Row sum of Squares + Error sum of Squares**

T.S.S. = R.S.S + E.S.S. DEGREES OF FREEDOM:

1. Degrees of freedom for Total Sum of Squares are **"N-1"**.

2. Degrees of freedom for Row Sum of Squares are **"k-1"**.
3. Degrees of freedom for Error Sum of Squares are **"N-k"**.

Where, N = Total number of observations, k = Number of rows.

MEAN SUM OF SQUARES (M.S.S.):

The sum of squares divide by its degrees of freedom gives the corresponding variance or Mean Sum of Squares.

1. Mean Sum of Squares due to Rows is: $\mathbf{S^2_R = R.S.S/(k-1).}$

2. Mean Sum of Squares due to Error is: $\mathbf{S^2_E = S.S.E/(N-k).}$ **TEST STATISTIC:**

Under Null Hypothesis (H_0), the test statistic is:

$\mathbf{F = [S^2_R / S^2_E] \sim F [\upsilon_1, \upsilon_2]}$ **d.f** (if $S^2 > S^2_E$) (OR) $\mathbf{F = [S^2_E / S^2_R] \sim F [\upsilon_2, \upsilon_1]}$ **d.f** (if $S^2_E > S$)

Where, $\mathbf{\upsilon_1 = k-1, \upsilon_2 = N-k}$

After calculating the calculated value, compare that value with table value. If the calculated value if "Less Than or Equal to (≤)" table value then accept the H_0 (Null Hypothesis) otherwise reject the H_0 (Null Hypothesis). $F_{cal} \leq F_{tab}$ (Accept H_0) $F_{cal} \geq F_{tab}$ (Reject H_0)

Source of Variation (S.V.)	Degrees of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares (M.S.S.)	Variance Ratio (F)
Due to Rows	k-1	S^2_R	S^2_R = R.S.S/(k-1)	$F_R = [S^2_R / S^2_E] \sim F [\upsilon_1, \upsilon_2]$ (OR) $F_E = [S^2_E / S^2_R] \sim F [\upsilon_2, \upsilon_1]$
Due to Error	N-k	S^2_E	S^2_E = S.S.E/(N-k)	
Total	N-1	S^2_T		

FORMULAE:

1. Total Sum of Squares = $S^2_T = \Sigma \Sigma x^2_{ij} - C.F.$
2. Row Sum of Squares = $S^2_R = \Sigma [T^2_i / n_i] - C.F.$
3. Error Sum of Squares = $S^2_E = S^2_T - S^2_R$

Where, G = Grand Total ($\Sigma \Sigma x_{ij}$), N = Number of observations,
C.F. Correction Factor = (G^2/N)

ANOVA TWO-WAY CLASSIFICATION:

Suppose the N observations are classified into „k" categories or classes according to some criteria
„A" and into „h" categories according to another criteria „B". This scheme of classification according to the two factors is called "ANOVA Two-Way Classification" and its analysis is "Two-Way Classification of ANOVA". In the two way classification, the values of the response variable are affected by two factors.

NULL HYPOTHESIS (H_0) AND ALTERNATIVE HYPOTHESIS (H_1):

FOR ROWS:

Null Hypothesis (H_{01}): There is no significant difference between means of „k" rows..

Therefore, $\mu_1 = \mu_2 = ... \mu_k = \mu$
Alternative Hypothesis (H_{11}): There is no significant difference between means of „k" rows.

FOR COLUMNS:

Null Hypothesis (H_{02}): There is no significant difference between means of „h" columns.

Therefore, $\mu_1 = \mu_2 = ... \mu_h = \mu$
Alternative Hypothesis (H_{12}): There is no significant difference between means of „h" columns.

STATISTICAL ANALYSIS:

Total Variation = Variation due to Rows + Variation due to Columns + Variation due to Error

Therefore, **T.S.S. = R.S.S. + C.S.S. + E.S.S.** **DEGREES OF FREEDOM:**

1. Degrees of freedom for Total Sum of Squares are **"N-1"**.

2. Degrees of freedom for Row Sum of Squares are **"k-1"**.
3. Degrees of freedom for Row Sum of Squares are **"h-1"**.
4. Degrees of freedom for Error Sum of Squares are **"(k-1) (h-1)"**.

Where, N = Total number of observations, k = Number of rows, h = Number of columns.

MEAN SUM OF SQUARES (M.S.S.):

The sum of squares divide by its degrees of freedom gives the corresponding variance or Mean Sum of Squares.

1. Mean Sum of Squares due to Rows is: $\mathbf{S^2_R = R.S.S/(k-1).}$
2. Mean Sum of Squares due to Columns is: $\mathbf{S^2_C = C.S.S/(h-1).}$
3. Mean Sum of Squares due to Error is: $\mathbf{S^2_E = S.S.E/[(k-1)\,(h-1)].}$ **TEST STATISTIC:**

Under Null Hypothesis (H_0), the test statistics are:

FOR ROWS:

$\mathbf{F = [S^2_R / S^2_E] \sim F\,[\upsilon_1, \upsilon_2]\ d.f_R}$ (if $S^2 > S^2_E$) (OR) $\mathbf{F = [S^2_E / S^2_R] \sim F\,[\upsilon_2, \upsilon_1]\ d.f_E}$ (if $S^2_E > S^2_R$)

Where, $\mathbf{\upsilon_1 = k-1,\ \upsilon_2 = (k-1)\,(h-1)}$

FOR COLUMNS:

$\mathbf{F = [S^2_C / S^2_E] \sim F\,[\upsilon_1, \upsilon_2]\ d.f_C}$ (if $S^2_C > S^2_E$) (OR) $\mathbf{F = [S^2_E / S^2_C] \sim F\,[\upsilon_2, \upsilon_1]\ d.f_E}$ (if $S^2_E > S^2_C$)

Where, $\mathbf{\upsilon_1 = h-1,\ \upsilon_2 = (k-1)\,(h-1)}$

After calculating the calculated value, compare that value with table value. If the calculated value if "Less Than or Equal to (≤)" table value then accept the H_0 (Null Hypothesis) otherwise reject the H_0 (Null Hypothesis). $F_{cal} \leq F_{tab}$ (Accept H_0) $F_{cal} \geq F_{tab}$ (Reject H_0)

Source of Variation (S.V.)	**Degrees of Freedom (d.f.)**	**Sum of Squares (S.S.)**	**Mean Sum of Squares (M.S.S.)**	**Variance Ratio (F)**
Due to Rows	k-1	S^2_R	S^2_R = R.S.S/(k-1)	**FOR ROWS:** $F = [S^2_R / S^2_E] \sim F\,[\upsilon_1, \upsilon_2]$ (OR) **FOR COLUMNS:** $F = [S^2_C / S^2_E] \sim F\,[\upsilon_1, \upsilon_2]$ (OR)
Due to Columns	h-1	S^2_C	S^2_C = C.S.S/(h-1)	
Due to Error	N-k	S^2_E	S^2_E = S.S.E/[(k-1) (h-1)]	
Total		S^2_T		

FORMULAE:

1. Total Sum of Squares = $S^2_T = \Sigma\Sigma x_{ij}^2 - C.F.$
2. Row Sum of Squares = $S^2_R = \sum_i [T_{i.}^2/n_i] - C.F.$
3. Column Sum of Squares = $S^2_C = \Sigma\,[T_{.j}^2/n_j] - C.F.$
4. Error Sum of Squares = $S^2_E = S^2_T - S^2_R - S^2_C$

Where, G = Grand Total ($\Sigma\Sigma x_{ij}$), N = Number of observations, C.F. Correction Factor = (G^2/N)

CHI-SQUARE (χ^2) TEST FOR GOODNESS OF FIT:

Suppose we are given a set of observed frequencies obtained from some experiments and we want to test whether these experimental observed or given values support a corresponding set of expected or theoretical frequencies. Karl Pearson developed a test for testing the significance of difference between observed and expected values. This test is popularly known as "Chi-Square test for Goodness of Fit" i.e. this test is used to test whether the difference between observed and expected frequencies are significant. The Test Statistic of χ^2 test is: $\chi^2 = \Sigma [(0_i - E_i)^2/E_i] \sim \chi^2_{n-1}$ **d.f. at α% Level of Significance**.

Where, 0_i = Observed values, E_i = Expected values, n-1 = Degrees of freedom

After calculating the calculated value it compares with a table value, if it is less than or equalsto accept H_0 otherwise reject H_0.

CONDITIONS FOR APPLYING χ^2 TEST:

The following conditions must be satisfied for the validity of χ^2 test of goodness of fit.

1. The sample observations should be independent.
2. Total frequencies (N) should be reasonably large, at least 50.
3. Sum of the observed frequencies is equal to the sum of the expected i.e. $\Sigma 0_i = \Sigma E_i = N$.
4. No observed cell frequency should be less than 5.
5. If it is the case, then it is combined with the preceding or succeeding frequency so thatthe combined frequency is more than 5.
6. This can be finally be adjusted for the degrees of freedom, lost in combination.

Unit-V Correlation and regression

INTRODUCTION

Statistical methods of measures of central tendency, dispersion, skewness and kurtosis are helpful for the purpose of comparison and analysis of distributions involving only one variable *i.e.* univariate distributions. However, describing the relationship between two or more variables, is another important part of statistics.

In many business research situations, the key to decision making lies in understanding the relationships between two or more variables. *For example,* in an effort to predict the behavior of the bond market, a broker might find it useful to know whether the interest rate of bonds is related to the prime interestrate. While studying the effect of advertising on sales, an account executive may find it useful to know whether there is a strong relationship between advertising dollars and sales dollars for a company.

The statistical methods of *Correlation* (discussed in the present lesson) and *Regression* (to be discussed in the next lesson) are helpful in knowing the relationship between two or more variables which may be related in same way, *like* interest rate of bonds and prime interest rate; advertising expenditure andsales; income and consumption; crop-yield and fertilizer used; height and weights and so on.

In all these cases involving two or more variables, we may be interested in seeing:

- if there is any association between the variables;
- if there is an association, is it strong enough to be useful;
- if so, what form the relationship between the two variables takes;
- how we can make use of that relationship for predictive purposes, that is, forecasting; and
- how good such predictions will be.

Since these issues are inter related, correlation and regression analysis, as two sides of a single process, consists of methods of examining the relationship between two or more variables. If two (or more) variables are correlated, we can use information about one (or more) variable(s) to predict the value of the other variable(s), and can measure the error of estimations - *a job of regression analysis.*

WHAT IS CORRELATION?

Correlation is a measure of association between two or more variables. When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in theother variable(s), they are said to be correlated.

—The correlation between variables is a measure of the nature and degree of association between the variables‖.

As *a measure of the degree of relatedness of two variables,* correlation is widely used in exploratory research when the objective is to locate variables that might be related in some way to the variable of interest.

TYPES OF CORRELATION

Correlation can be classified in several ways. The important ways of classifying correlation are:

(i) Positive and negative,

(ii) Linear and non-linear (curvilinear) and

(iii) Simple, partial and multiple.

Positive and Negative Correlation

If both the variables move in the same direction, we say that there is a positive correlation, *i.e.,* if one variable increases, the other variable also increases on an average or if one variable decreases, the other variable also decreases on an average.

On the other hand, if the variables are varying in opposite direction, we say that it is a case of negative correlation; *e.g.,* movements of demand and supply.

Linear and Non-linear (Curvilinear) Correlation

If the change in one variable is accompanied by change in another variable in a constant ratio, it is acase of linear correlation. Observe the following data:

X	:	10	20	30	40	50
Y	:	25	50	75	100	125

The ratio of change in the above example is the same. It is, thus, a case of linear correlation. If we plot these variables on graph paper, all the points will fall on the same straight line.

On the other hand, if the amount of change in one variable does not follow a constant ratio with the change in another variable, it is a case of non-linear or curvilinear correlation. If a couple of figures in either series *X* or series *Y* are changed, it would give a non-linear correlation.

Simple, Partial and Multiple Correlation

The distinction amongst these three types of correlation depends upon the number of variables involved in a study. If only two variables are involved in a study, then the correlation is said to be simple correlation. When three or more variables are involved in a study, then it is a problem of either partial or multiple correlation. In multiple correlation, three or more variables are studied simultaneously. But in partial correlation we consider only two variables influencing each other while the effect of other variable(s) is held constant.

Suppose we have a problem comprising three variables *X, Y* and *Z. X* is the number of hours studied, *Y*is I.Q. and *Z* is the number of marks obtained in the examination. In a multiple correlation, we will study the relationship between the marks obtained *(Z)* and the two variables, number of hours studied *(X)* and I.Q. (*Y*). In contrast, when we study the relationship between *X* and *Z*, keeping an average I.Q. *(Y)* as constant, it is said to be a study involving partial correlation. In this lesson, we will study linear correlation between two variables.

CORRELATION DOES NOT NECESSARILY MEAN CAUSATION

The correlation analysis, in discovering the nature and degree of relationship between variables, does not necessarily imply any cause and effect relationship between the variables. Two variables may be related to each other but this does not mean that one variable causes the other. *For example,* we may find that logical reasoning and creativity are correlated, but that does not mean if we could increase peoples' logical reasoning ability, we would produce greater creativity. We need to conduct an actual experiment to unequivocally demonstrate a causal relationship. But if it is true that influencing someones' logical reasoning ability does influence their creativity, then the two variables must be correlated with each other. In other words, *causation always implies correlation, however converse is not true.* Let us see some situations:

1. The correlation may be due to chance particularly when the data pertain to a small sample. A small sample bivariate series may show the relationship but such a relationship may not exist in the universe.
2. It is possible that both the variables are influenced by one or more other variables. For example, expenditure on food and entertainment for a given number of households show a positive relationship because both have increased over time. But, this is due to rise in family incomes over the same period. In other words, the two variables have been influenced by another variable - increase in family incomes.
3. There may be another situation where both the variables may be influencing each other so that we cannot say which is the cause and which is the effect. *For example,* take the case of price and demand. The rise in price of a commodity may lead to a decline in the demand for it. Here, price is the cause and the demand is the effect. In yet another situation, an increase in demand may lead to a

rise in price. Here, the demand is the cause while price is the effect, which is just the reverse of the earlier situation. In such situations, it is difficult to identify which variable is causing the effect on which variable, as both are influencing each other.

The foregoing discussion clearly shows that correlation does not indicate any causation or functional relationship. Correlation coefficient is merely a mathematical relationship and this has nothing to do with cause and effect relation. It only reveals co-variation between two variables. Even when there is nocause-and-effect relationship in bivariate series and one interprets the relationship as causal, such a correlation is called spurious or non-sense correlation. Obviously, this will be misleading. As such, one has to be very careful in correlation exercises and look into other relevant factors before concluding a cause-and-effect relationship.

CORRELATION ANALYSIS

Correlation Analysis is a statistical technique used to indicate the nature and degree of relationship existing between one variable and the other(s). It is also used along with regression analysis to measure how well the regression line explains the variations of the dependent variable with the independent variable.

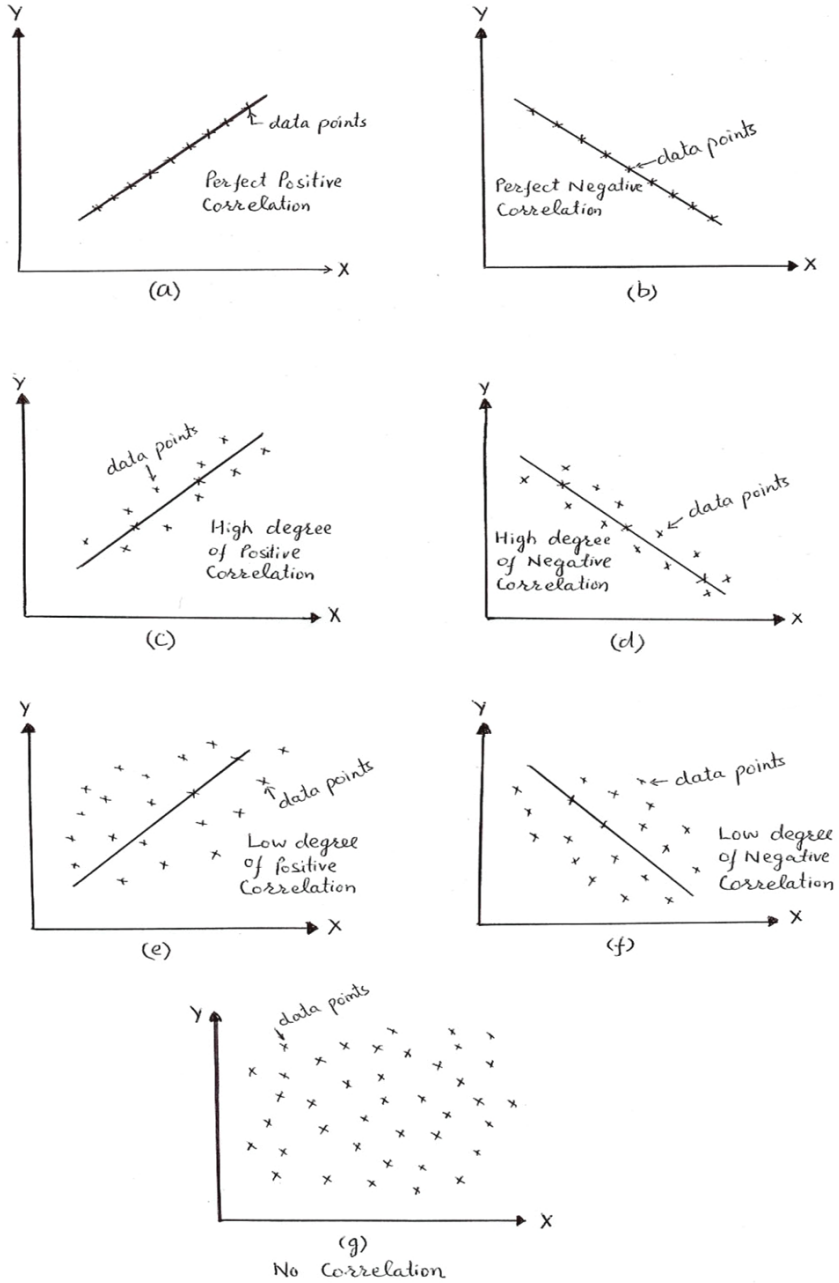
The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include:

1. Scatter Diagram
2. Correlation Graph
3. Pearson‘s Coefficient of Correlation
4. Spearman‘s Rank Correlation
5. Concurrent Deviation Method

SCATTER DIAGRAM

This method is also known as Dotogram or Dot diagram. Scatter diagram is one of the simplest methods of diagrammatic representation of a bivariate distribution. Under this method, both the variables are plotted on the graph paper by putting dots. The diagram so obtained is called "Scatter Diagram". By

studying diagram, we can have rough idea about the nature and degree of relationship between two variables. The term scatter refers to the spreading of dots on the graph. We should keep the following points in mind while interpreting correlation:



- if the plotted points are very close to each other, it indicates high degree of correlation. If the plotted points are away from each other, it indicates low degree of correlation.
- if the points on the diagram reveal any trend (either upward or downward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.
- if there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner, the correlation is positive since this reveals that the values of the two variables move in

the same direction. If, on the other hand, the points depict a downward trend from the upper left hand corner to the lower right hand corner, the correlation is negative since in this case the values of the two variables move in the opposite directions.

- in particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points like on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative.

The various diagrams of the scattered data in Figure 12.1 depict different forms of correlation.

Example 12-1

Given the following data on sales (in thousand units) and expenses (in thousand rupees) of a firm for 10month:

Month :	J	F	M	A	M	J	J	A	S	O
Sales:	50	50	55	60	62	65	68	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

a) Make a Scatter Diagram

b) Do you think that there is a correlation between sales and expenses of the firm? Is it positive or negative? Is it high or low?

Solution:(a) The Scatter Diagram of the given data is shown in Figure 4-2

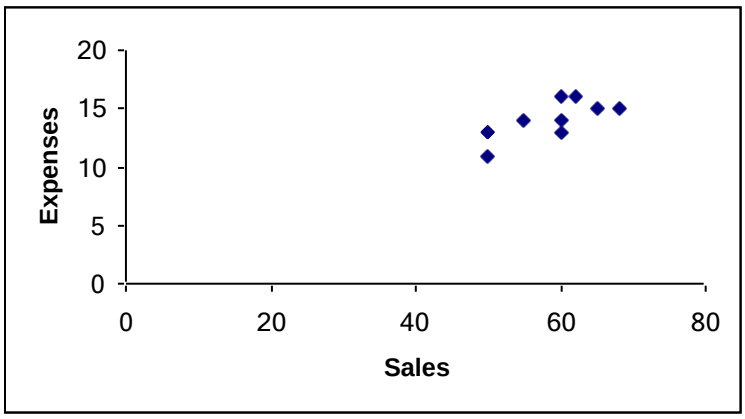


Figure 12.2 Scatter Diagram

(b) Figure 12.2 shows that the plotted points are close to each other and reveal an upward trend. Sothere is a high degree of positive correlation between sales and expenses of the firm.

CORRELATION GRAPH

This method, also known as Correlogram is very simple. The data pertaining to two series are plotted ona graph sheet. We can find out the correlation by examining the direction and closeness of two curves.If both the curves drawn on the graph are moving in the same direction, it is a case of positive correlation. On the other hand, if both the curves are moving in opposite direction, correlation is said to be negative. If the graph does not show any definite pattern on account of erratic fluctuations in the curves, then it shows an absence of correlation.

Example 12.2

Find out graphically, if there is any correlation between price yield per plot (qtls); denoted by *Y* andquantity of fertilizer used (kg); denote by *X*.

Plot No.:	1	2	3	4	5	6	7	8	9	10
Y:	3.5	4.3	5.2	5.8	6.4	7.3	7.2	7.5	7.8	8.3
X:	6	8	9	12	10	15	17	20	18	24

Solution: The Correlogram of the given data is shown in Figure 4-3

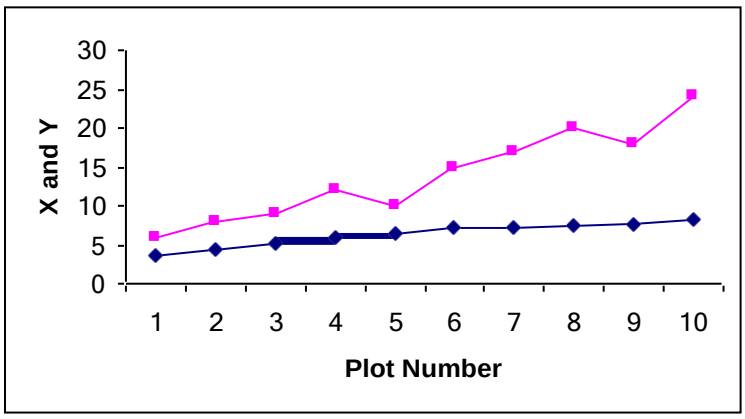


Figure 12.3 Correlation Graph

Figure 12.3 shows that the two curves move in the same direction and, moreover, they are very close to each other, suggesting a close relationship between price yield per plot (qtls) and quantity of fertilizer used (kg)

Remark: Both the Graphic methods - scatter diagram and correlation graph provide a *„feel for'* of the data – by providing visual representation of the association between the variables. These are readily comprehensible and enable us to form a fairly good, though rough idea of the nature and degree of the relationship between the two variables. However, these methods are unable to quantify the relationship between them. To quantify the extent of correlation, we make use of algebraic methods - which calculate correlation coefficient.

PEARSON'S COEFFICIENT OF CORRELATION

A mathematical method for measuring the intensity or the magnitude of *linear relationship* between twovariables was suggested by Karl Pearson (1867-1936), a great British Biometrician and Statistician and, it is by far the most widely used method in practice.

Karl Pearson's measure, known as Pearsonian correlation coefficient between two variables *X* and *Y*, usually denoted by *r(X,Y)* or r_{xy} or simply *r* is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between *X* and *Y*, to the product of the standard deviationsof *X* and *Y*.

Symbolically

$$r_{xy} = \frac{Cov(X,Y)}{S_x.S_y} \qquad \ldots\ldots\ldots\ldots(4.1)$$

when, $(X_1,Y_1);(X_2,Y_2);\ .....(X_n,Y_n)$ are *N* pairs of observations of the variables *X* and *Y* in a bivariate distribution,

$$Cov(X,Y)=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{N} \qquad \ldots\ldots\ldots(4.2a)$$

$$S_x=\sqrt{\frac{\sum(X-\overline{X})^2}{N}} \qquad \ldots\ldots\ldots(4.2b)$$

$$\text{and } S_y=\sqrt{\frac{\sum(Y-\overline{Y})^2}{N}} \qquad \ldots\ldots\ldots(4.2c)$$

Thus by substituting *Eqs. (4.2)* in *Eq. (4.1),* we can write the Pearsonian correlation coefficient as

$$r_{xy}=\frac{\frac{1}{N}\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\frac{1}{N}\sum(X-\overline{X})^2}\sqrt{\frac{1}{N}\sum(Y-\overline{Y})^2}}$$

$$r_{xy}=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}} \qquad \ldots\ldots\ldots(4.3)$$

If we denote, $d_x=X-\overline{X}$ and $d_y=Y-\overline{Y}$

Then $$r_{xy}=\frac{\sum d_x d_y}{\sqrt{\sum d_x{}^2 d_y{}^2}} \qquad \ldots\ldots\ldots(4.3a)$$

We can further simply the calculations of *Eqs. (4.2)*

We have

$$Cov(X,Y)=\frac{1}{N}\sum(X-\overline{X})(Y-\overline{Y})$$

$$=\frac{1}{N}\sum XY-\overline{X}\overline{Y}$$

$$=\frac{1}{N}\sum XY-\frac{\sum X}{N}\frac{\sum Y}{N}$$

$$=\frac{1}{N^2}\left[N\sum XY-\sum X\sum Y\right] \qquad \ldots\ldots\ldots(4.4)$$

and $S_x^2 = \frac{1}{N}\sum(X-\bar{X})^2$

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

$$= \frac{1}{N}\sum X^2 - \left(\frac{\sum X}{N}\right)^2$$

$$= \frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right] \qquad \ldots\ldots\ldots(4.5a)$$

Similarly, we have

$$S_y^2 = \frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right] \qquad \ldots\ldots\ldots(4.5b)$$

So Pearsonian correlation coefficient may be found as

$$r_{xy} = \frac{\frac{1}{N^2}\left[N\sum XY - \sum X\sum Y\right]}{\sqrt{\frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right]}\sqrt{\frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

or

$$r_{xy} = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \qquad \ldots\ldots\ldots(4.6)$$

Remark: *Eq. (4.3)* or *Eq. (4.3a)* is quite convenient to apply if the means $\bar{X}$ and $\bar{Y}$ come out to be integers. If $\bar{X}$ or/and $\bar{Y}$ is (are) fractional then the *Eq. (4.3)* or *Eq. (4.3a)* is quite cumbersome to apply, since the computations of $\sum(X-\bar{X})^2$, $\sum(Y-\bar{Y})^2$ and $\sum(X-\bar{X})(Y-\bar{Y})$ are quite time consuming and tedious. In such a case *Eq. (4.6)* may be used provided the values of *X* or/ and *Y* are small. But if *X* and *Y* assume large values, the calculation of $\sum X^2$, $\sum Y^2$ and $\sum XY$ is again quite time consuming. Thus if *(i)* *X* and *Y* are fractional and *(ii)* *X* and *Y* assume large values, the *Eq. (4.3)* and *Eq. (4.6)* arenot generally used for numerical problems. In such cases, the step deviation method where we take the deviations of the variables *X* and *Y* from any arbitrary points is used. We will discuss this method in theproperties of correlation coefficient.

Properties of Pearson Correlation Coefficient

The following are important properties of Pearson correlation coefficient:

1. *Pearson correlation coefficient cannot exceed 1 numerically*. In other words it lies between –1 and +1. Symbolically, $-1 \leq r \leq 1$

Remarks:

(i) This property provides us a check on our calculations. If in any problem, the obtained value of *r* lies outside the limits $\pm$ 1, this implies that there is some mistake in our calculations.

(ii) The sign of *r* indicate the nature of the correlation. Positive value of *r* indicates positive correlation, whereas negative value indicates negative correlation. *r = 0* indicate absence of correlation.

(iii) The following table sums up the degrees of correlation corresponding to various values of *r:*

Value of *r*	**Degree of correlation**
±1	Perfect correlation
±0.90 or more	Very high degree of correlation
±0.75 to ±0.90	Sufficiently high degree of correlation
±0.60 to ±0.75	Moderate degree of correlation
±0.30 to ±0.60	Only the possibility of a correlation
less than ±0.30	Possibly no correlation
0	Absence of correlation

2. *Pearsonian Correlation coefficient is independent of the change of origin and scale*. Mathematically, if given variables *X* and *Y* are transformed to new variables *U* and *V* by change of origin and scale, *i. e.*

$$U = \frac{X - A}{h} \quad \text{and} \quad V = \frac{Y - B}{k}$$

Where *A, B, h* and *k* are constants and *h > 0, k > 0;* then the correlation coefficient between *X* and *Y* is same as the correlation coefficient between *U* and *V i.e.*,

$$r(X,Y) = r(U, V) \Rightarrow r_{xy} = r_{uv}$$

Remark: This is one of the very important properties of the correlation coefficient and is extremelyhelpful in numerical computation of *r*. We had already stated that *Eq. (4.3)* and *Eq.(4.6)* become quite

tedious to use in numerical problems if X and/or Y are in fractions or if X and Y are large. In such cases we can conveniently change the origin and scale (if possible) in X or/and Y to get new variables U and $V$and compute the correlation between U and V by the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \sum U \sum V}{\sqrt{N\sum U^2 - \left(\sum U\right)^2}\sqrt{N\sum V^2 - (V)^2}} \quad \ldots\ldots\ldots\ldots (4.7)$$

3. Two independent variables are uncorrelated but the converse is not true

If X and Y are independent variables then

$$r_{xy} = 0$$

However, the converse of the theorem is not true *i.e.,* uncorrelated variables need not necessarily be independent. As an illustration consider the following bivariate distribution.

X	:	1	2	3	-3	-2	-1
Y	:	1	4	9	9	4	1

For this distribution, value of r will be 0.

Hence in the above example the variable X and Y are uncorrelated. But if we examine the data carefully we find that X and Y are not independent but are connected by the relation $Y = X^2$. The above example illustrates that uncorrelated variables need not be independent.

Remarks: One should not be confused with the words uncorrelation and independence. $r_{xy} = 0$ *i.e.,* uncorrelation between the variables X and Y simply implies the absence of any linear (straight line) relationship between them. They may, however, be related in some other form other than straight line *e.g.,* quadratic (as we have seen in the above example), logarithmic or trigonometric form.

4. *Pearson coefficient of correlation is the geometric mean of the two regression coefficients, i.e.*

$$r_{xy} = \pm\sqrt{b_{xy}.b_{yx}}$$

The signs of both the regression coefficients are the same, and so the value of r will also have the samesign.

This property will be dealt with in detail in the next lesson on Regression Analysis.

5. *The square of Pearsonian correlation coefficient is known as the coefficient of determination*.

Coefficient of determination, which measures the percentage variation in the dependent variable that is accounted for by the independent variable, is a much better and useful measure for interpreting the value of r. This property will also be dealt with in detail in the next lesson.

Probable Error of Correlation Coefficient

The correlation coefficient establishes the relationship of the two variables. After ascertaining this level of relationship, we may be interested to find the extent upto which this coefficient is dependable. Probable error of the correlation coefficient is such a measure of testing the reliability of the observed value of the correlation coefficient, when we consider it as satisfying the conditions of the random sampling.

If *r* is the observed value of the correlation coefficient in a sample of *N* pairs of observations for the twovariables under consideration, then the Probable Error, denoted by *PE* (*r*) is expressed as

$$PE(r) = 0.6745\ SE(r)$$

or

$$PE(r) = 0.6745\frac{1-r^2}{\sqrt{N}}$$

There are two main functions of probable error:

1. **Determination of limits**: The limits of population correlation coefficient are $r \pm PE(r)$, implying that if we take another random sample of the size *N* from the same population, then the observed value of the correlation coefficient in the second sample can be expected to lie within the limits given above, with 0.5 probability. When sample size *N* is small, the concept or value of *PE* may lead to wrong conclusions. Hence to use the concept of *PE* effectively, sample size *N* it should be fairly large.
2. ***Interpretation of 'r':*** The interpretation of *'r'* based on *PE* is as under:
   - If $r < PE(r)$, there is no evidence of correlation, *i.e.* a case of insignificant correlation.
   - If $r > 6\ PE(r)$, correlation is significant. *If* $r < 6\ PE(r)$, it is insignificant.
   - If the probable error is small, correlation exist where $r > 0.5$

Example 12-3

Find the Pearsonian correlation coefficient between sales (in thousand units) and expenses (in thousandrupees) of the following 10 firms:

Firm:	1	2	3	4	5	6	7	8	9	10
Sales:	50	50	55	60	65	65	65	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

Solution: Let sales of a firm be denoted by *X* and expenses be denoted by *Y*

Calculations for Coefficient of Correlation

{Using *Eq. (4.3)* or *(4.3a)*}

Firm	X	Y	$d_x = X - \bar{X}$	$d_y = Y - \bar{Y}$	d_x^2	$_y d^2$	$d_x.d_y$
1	50	11	-8	-3	64	9	24
2	50	13	-8	-1	64	1	8
3	55	14	-3	0	9	0	0
4	60	16	2	2	4	4	4
5	65	16	7	2	49	4	14
6	65	15	7	1	49	1	7
7	65	15	7	1	49	1	7
8	60	14	2	0	4	0	0
9	60	13	2	-1	4	1	-2
10	50	13	-8	-1	64	1	8
	$\sum X$ = 580	$\sum Y$ = 140			$\sum d_x^2$ =360	$\sum d_y^2$ =22	$\sum d_x d_y$ =70

$$\bar{X} = \frac{\sum X}{N} = \frac{580}{10} = 58 \quad \text{and} \quad \bar{Y} = \frac{\sum Y}{N} = \frac{140}{10} = 14$$

Applying the *Eq. (4.3a), w*e have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^{\,2} d_y^{\,2}}}$$

$$r_{xy} = \frac{70}{\sqrt{360x22}}$$

$$r_{xy} = \frac{70}{\sqrt{7920}} = 0.78$$

The value of $r_{xy} = 0.78$, indicate a high degree of positive correlation between sales and expenses.

Example 12-4

The data on price and quantity purchased relating to a commodity for 5 months is given below:

Month :	January	February	March	April	May
Prices(Rs):	10	10	11	12	12
Quantity(Kg):	5	6	4	3	3

Find the Pearsonian correlation coefficient between prices and quantity and comment on its sign andmagnitude.

Solution: Let price of the commodity be denoted by *X* and quantity be denoted by *Y*

Calculations for Coefficient of Correlation

{Using *Eq. (4.6)*}

Month	X	Y	X^2	Y^2	XY
1	10	5	100	25	50
2	10	6	100	36	60
3	11	4	121	16	44
4	12	3	144	9	36
5	12	3	144	9	36
	$\sum X = 55$	$\sum Y = 21$	$\sum X^2 = 609$	$\sum Y^2 = 95$	$\sum XY = 226$

Applying the *Eq. (4.6), w*e have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r_{xy} = \frac{5x226 - 55x21}{\sqrt{(5x609 - 55x55)(5x95 - 21x21)}}$$

$$r_{xy} = \frac{1130 - 1155}{\sqrt{20x34}} =$$

$$r_{xy} = \frac{-25}{\sqrt{680}}$$

$$r_{xy} = -0.98$$

The negative sign of *r* indicate negative correlation and its large magnitude indicate a very high degreeof correlation. So there is a high degree of negative correlation between prices and quantity demanded. ***Example 12-5***

Find the Pearsonian correlation coefficient from the following series of marks obtained by 10 studentsin a class test in mathematics (*X)* and in Statistics (*Y)*:

X :	45	70	65	30	90	40	50	75	85	60
Y :	35	90	70	40	95	40	60	80	80	50

Also calculate the Probable Error.

Solution:

Calculations for Coefficient of Correlation

{Using *Eq. (4.7)*}

X	*Y*	*U*	*V*	U^2	V^2	*UV*
45	35	-3	-6	9	36	18
70	90	2	5	4	25	10
65	70	1	1	1	1	1
30	40	-6	-5	36	25	30
90	95	6	6	36	36	36
40	40	-4	-5	16	25	20

50	60	-2	-1	4	1	2
75	80	3	3	9	9	9
85	80	5	3	25	9	15
60	50	0	-3	0	9	0
		$\sum U = 2$	$\sum V = -2$	$\sum U^2 = 140$	$\sum V^2 = 176$	$\sum UV = 141$

We have, defined variables *U* and *V* as

$$U = \frac{X - 60}{5} \qquad \text{and} \qquad V = \frac{Y - 65}{5}$$

Applying the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \left(\sum U \sum V\right)}{\sqrt{N\sum U^2 - \left(\sum U\right)^2}\sqrt{N\sum V^2 - \left(\sum V\right)^2}}$$

$$= \frac{10x141 - 2x(-2)}{\sqrt{10x140 - 2x2}\sqrt{10x176 - (-2)x(-2)}}$$

$$= \frac{1410 + 4}{\sqrt{1400 - 4}\sqrt{1760 - 4}}$$

$$= \frac{1414}{\sqrt{2451376}} = 0.9$$

So, there is a high degree of positive correlation between marks obtained in Mathematics and inStatistics. Probable Error, denoted by *PE* (*r*) is given as

$$PE(r) = 0.6745\frac{1 - r^2}{\sqrt{N}}$$

$$PE(r) = 0.6745\frac{1 - (0.9)^2}{\sqrt{10}}$$

$$PE(r) = 0.0405$$

So the value of *r* is highly significant.

SPEARMAN'S RANK CORRELATION

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, *etc.*, which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904, which consists in obtaining the correlation coefficient between the ranks of *N* individuals in the two attributes under study.

Suppose we want to find if two characteristics *A*, say, intelligence and *B*, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of *N* individuals in order of merit (ranks) *w.r.t.* proficiency in the two characteristics. Let the random variables *X* and *Y* denote the ranks of the individuals in the characteristics *A* and *B* respectively. If we assume that there is no tie, *i.e.*, if no two individuals get the same rank in a characteristic then, obviously, *X* and *Y* assume numerical values ranging from *1* to *N*.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for the group of individuals.

Spearman's rank correlation coefficient, usually denoted by ρ(Rho) is given by the equation

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} \qquad \ldots\ldots\ldots\ldots (4.8)$$

Where *d* is the difference between the pair of ranks of the same individual in the two characteristics and *N* is the number of pairs.

Example 12-6

Ten entries are submitted for a competition. Three judges study each entry and list the ten in rank order.Their rankings are as follows:

Entry:	A	B	C	D	E	F	G	H	I	J
Judge J_1:	9	3	7	5	1	6	2	4	10	8
Judge J_2:	9	1	10	4	3	8	5	2	7	6
Judge J_3:	6	3	8	7	2	4	1	5	9	10

Calculate the appropriate rank correlation to help you answer the following questions:

(i) Which pair of judges agrees the most?

(ii) Which pair of judges disagrees the most?

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq.(4.8)*}

Entry	Rank by Judges			Difference in Ranks					
	J_1	J_2	J_3	$d(J_1 \& J_2)$	d^2	$d(J_1 \& J_3)$	d^2	$d(J_2 \& J_3)$	d^2
A	9	9	6	0	0	+3	9	+3	9
B	3	1	3	+2	4	0	0	-2	4
C	7	10	8	-3	9	-1	1	+2	4
D	5	4	7	+1	1	-2	4	-3	9
E	1	3	2	-2	4	-1	1	+1	1
F	6	8	4	-2	4	+2	4	+4	16
G	2	5	1	-3	9	+1	1	+4	16
H	4	2	5	+2	4	-1	1	-3	9
I	10	7	9	+3	9	+1	1	-2	4
J	8	6	10	+2	4	-2	4	-4	16
					$\sum d^2 = 48$		$\sum d^2 = 26$		$\sum d^2 = 88$

$$\rho\,(J_1 \& J_2) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 48}{10(10^2-1)} = 1 - \frac{288}{990} = 1 - 0.29 = +0.71$$

$$\rho\,(J_1 \& J_3) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 26}{10(10^2-1)} = 1 - \frac{156}{990} = 1 - 0.1575 = +0.8425$$

$$\rho\,(J_2 \& J_3) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 88}{10(10^2-1)} = 1 - \frac{528}{990} = 1 - 0.53 = +0.47$$

So *(i)* Judges J_1 and J_3 agree the most

(ii) Judges J_2 and J_3 disagree the most

Spearman's rank correlation *Eq.(4.8)* can also be used even if we are dealing with variables, which are

measured quantitatively, *i.e.* when the actual data but not the ranks relating to two variables are given. In such a case we shall have to convert the data into ranks. The highest (or the smallest) observation is given the rank 1. The next highest (or the next lowest) observation is given rank 2 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for all the variables under consideration.

Example 12-7

Calculate the rank coefficient of correlation from the following data:

X:	75	88	95	70	60	80	81	50
Y:	120	134	150	115	110	140	142	100

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq.(4.8)*}

X	Ranks R_X	Y	Ranks R_Y	$d = R_X - R_Y$	d^2
75	5	120	5	0	0
88	2	134	4	-2	4
95	1	150	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
80	4	140	3	+1	1
81	3	142	2	+1	1
50	8	100	8	0	0

$\sum d^2 = 6$

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - \frac{6 \times 6}{8(8^2-1)} = 1 - \frac{36}{504} = 1 - 0.07 = +0.93$$

Hence, there is a high degree of positive correlation between *X* and *Y*

Repeated Ranks

In case of attributes if there is a tie *i.e.,* if any two or more individuals are placed together in anyclassification *w.r.t.* an attribute or if in case of variable data there is more than one item with the same

value in either or both the series then Spearman's *Eq.(4.8)* for calculating the rank correlation coefficient breaks down, since in this case the variables *X* [the ranks of individuals in characteristic A (1st series)] and *Y* [the ranks of individuals in characteristic B (2nd series)] do not take the values from *1* to *N*.

In this case common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks, which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common rank. For example, suppose an item is repeated at rank 4. Then the common rank to be assigned to each item is (4+5)/2, *i.e.,* 4.5 which is the average of 4 and 5, the ranks which these observations would have assumed if they were different. The next item will be assigned the rank 6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will be (7+8+9)/3, *i.e.,* 8 which is the arithmetic mean of 7,8 and 9 *viz.,* the ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

If only a small proportion of the ranks are tied, this technique may be applied together with *Eq.(4.8)*. If a large proportion of ranks are tied, it is advisable to apply an adjustment or a correction factor to *Eq.(4.8)* as explained below:

—In the Eq.(4.8) add the factor

$$\frac{m(m^2-1)}{12} \qquad \ldots\ldots\ldots(4.8a)$$

to $\sum d^2$*; where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the series‖.*

Example 12-8

For a certain joint stock company, the prices of preference shares (*X*) and debentures (*Y*) are given below:

X:	73.2	85.8	78.9	75.8	77.2	81.2	83.8
Y:	97.8	99.2	98.8	98.3	98.3	96.7	97.1

Use the method of rank correlation to determine the relationship between preference prices and debentures prices.

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq. (4.8)* and *(4.8a)*}

X	Y	Rank of X (X_R)	Rank of Y (Y_R)	$d = X_R - Y_R$	d^2
73.2	97.8	7	5	2	4
85.8	99.2	1	1	0	0
78.9	98.8	4	2	2	4
75.8	98.3	6	3.5	2.5	6.25
77.2	98.3	5	3.5	1.5	2.25
81.2	96.7	3	7	-4	16
83.8	97.1	2	6	-4	16
				$\sum d = 0$	$\sum d^2 = 48.50$

In this case, due to repeated values of Y, we have to apply ranking as average of 2 ranks, which couldhave been allotted, if they were different values. Thus ranks 3 and 4 have been allotted as 3.5 to both the values of Y = 98.3. Now we also have to apply correction factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m in the number of times the value is repeated, here $m = 2$.

$$\rho = \frac{6\left[\sum d^2 + \frac{m(m^2-1)}{2}\right]}{N(N^2-1)} = \frac{6\left[48.5 + \frac{2(4-1)}{12}\right]}{7(7^2-1)} = 1 - \frac{6 \times 49}{7 \times 48} = 0.125$$

Hence, there is a very low degree of positive correlation, probably no correlation, between preference share prices and debenture prices.

Remarks on Spearman's Rank Correlation Coefficient

1. We always have $\sum d = 0$, which provides a check for numerical calculations.
2. Since Spearman's rank correlation coefficient, ρ, is nothing but Karl Pearson's correlation coefficient, r, between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.
3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure, which is distribution free (or non-parametric). Spearman's ρ is such a distribution free measure, since no strict assumption are made about the from of the population from which sample observations are drawn.
4. Spearman's formula is easy to understand and apply as compared to Karl Pearson's formula. The values obtained by the two formulae, *viz* Pearsonian r and Spearman's ρ are generally different. The

difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics, which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.
6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution. For $N > 30$, this formula should not be used unless the ranks are given.

CONCURRENT DEVIATION METHOD

This is a casual method of determining the correlation between two series when we are not very serious about its precision. This is based on the signs of the deviations (*i.e.* the direction of the change) of the values of the variable from its preceding value and does not take into account the exact magnitude of the values of the variables. Thus we put a plus (+) sign, minus (-) sign or equality (=) sign for the deviation if the value of the variable is greater than, less than or equal to the preceding value respectively. The deviations in the values of two variables are said to be concurrent if they have the same sign (either both deviations are positive or both are negative or both are equal). The formula used for computing correlation coefficient r_c by this method is given by

$$r_c = \pm\sqrt{\pm\left(\frac{2c-N}{N}\right)} \quad \ldots\ldots\ldots\ldots (4.9)$$

Where c is the number of pairs of concurrent deviations and N is the number of pairs of deviations. If $(2c-N)$ is positive, we take positive sign in and outside the square root in *Eq. (4.9)* and if $(2c-N)$ is negative, we take negative sign in and outside the square root in *Eq. (4.9).*

Remarks: (i) It should be clearly noted that here N is not the number of pairs of observations but it is the number of pairs of deviations and as such it is one less than the number of pairs of observations.

(ii) Coefficient of concurrent deviations is primarily based on the following principle:

> *—If the short time fluctuations of the time series are positively correlated or in other words, if their deviations are concurrent, their curves would move in the same direction and would indicate positive correlation between them‖*

1.Calculate coefficient of correlation by the concurrent deviation method

Supply:	112	125	126	118	118	121	125	125	131	135
Price:	106	102	102	104	98	96	97	97	95	90

Calculations for Coefficient of Concurrent Deviations

Supply (X)	Sign of deviation from preceding value (X)	Price (Y)	Sign of deviation preceding value (Y)	Concurrent deviations
112		106		
125	+	102	-	
126	+	102	=	
118	-	104	+	
118	=	98	-	
121	+	96	-	
125	+	97	+	+(c)
125	=	97	=	= (c)
131	+	95	-	
135	+	90	-	

We have

Number of pairs of deviations, N =10 – 1 = 9

c = Number of concurrent deviations

= Number of deviations having like signs

= 2

Coefficient of correlation by the method of concurrent deviations is given by:

$$r_c = \pm\sqrt{\pm\left|\frac{2c-N}{N}\right|}$$

$$r_c = \pm\sqrt{\pm\left|\frac{2x2-9}{9}\right|}$$

$$r_c = \pm\sqrt{\pm(-0.5556)}$$

Since $2c - N$ = -5 (negative), we take negative sign inside and outside the square root

$$r_c = -\sqrt{-(-0.5556)}$$

$$r_c = -\sqrt{0.5556}$$

$$r_c = -0.7$$

Hence there is a fairly good degree of negative correlation between supply and price.

LIMITATIONS OF CORRELATION ANALYSIS

As mentioned earlier, correlation analysis is a statistical tool, which should be properly used so that correct results can be obtained. Sometimes, it is indiscriminately used by management, resulting in misleading conclusions. We give below some *errors* frequently made in the use of correlation analysis:

1. Correlation analysis cannot determine cause-and-effect relationship. One should not assume that a change in *Y* variable is caused by a change in *X* variable unless one is reasonably sure that one variable is the cause while the other is the effect. Let us take an example.

   Suppose that we study the performance of students in their graduate examination and their earnings after, say, three years of their graduation. We may find that these two variables are highly and positively related. At the same time, we must not forget that both the variables might have been influenced by some other factors such as quality of teachers, economic and social status of parents, effectiveness of the interviewing process and so forth. If the data on these factors are available, then it is worthwhile to use multiple correlation analysis instead of bivariate one.

2. Another mistake that occurs frequently is on account of misinterpretation of the coefficient of correlation. Suppose in one case $r = 0.7$, it will be wrong to interpret that correlation explains 70 percent of the total variation in *Y*. The error can be seen easily when we calculate the coefficient of determination. Here, the coefficient of determination r^2 will be 0.49. This means that only 49 percent of the total variation in *Y* is explained.

   Similarly, the coefficient of determination is misinterpreted if it is also used to indicate causal relationship, that is, the percentage of the change in one variable is due to the change in another variable.

3. Another mistake in the interpretation of the coefficient of correlation occurs when one concludes a positive or negative relationship even though the two variables are actually unrelated. For example, the age of students and their score in the examination have no relation with each other. The two variables may show similar movements but there does not seem to be a common link between them.

To sum up, one has to be extremely careful while interpreting coefficient of correlation. Before one concludes a causal relationship, one has to consider other relevant factors that might have any influence

on the dependent variable or on both the variables. Such an approach will avoid many of the pitfalls in the interpretation of the coefficient of correlation. It has been rightly said that the coefficient ofcorrelation is not only one of the most widely used, but also one of the widely abused statistical measures.

PARTIAL CORRELATION

It is a statistical technique to analyze the association between dependent variables and one of the independent variables by eliminating the effect of other variables. Partial correlation is also known as Net Correlation. It is a statistical technique to study the association between one dependent variable and one independent variable by keeping other independent variables constant. In simple correlation, the effect of other independent variables was ignored.