

# 1. INTRODUCTION

## 1.1 Historical Development of Statistics

The word "Statistics" seems to have obtained from the Latin word "Status" or the Italian word "Statista" or the German word "Statistik" each of which means "Political State". In ancient time, the government used to collect informations about total population, land, wealth, total number of employees, soldiers etc. to have the idea of the manpower of the country for formulation of administrative set-up, fiscal, new taxes, levies and military policies of the government.

More than 2000 years ago, Chandra Gupta Maurya (324-300 B. C.) made arrangement of collecting official and administrative statistics and a good collection of vital statistics and registration of births and deaths were mentioned in Kautilya's Arthashastra which was published even before 300 B. C. During Akbar's reign, Abul Fazal wrote 'Ain-i-Akbari' in which a good account of population and statistical survey was given.

In mid-seventeenth century, the theoretical development in modern statistics came with the introduction of "Theory of Probability" and "Theory of Games and Chances". The French mathematician Pascal (1623-1662) solved the famous 'Problem of Points' which laid the foundation of the modern theory of probability. In this field other important contributor is James Bernoulli (1654-1705) who wrote the first treatise on the "Theory of Probability." De Moivre (1667-1754) and Laplace (1749-1820) also worked on the theory of probability. Gauss (1777-1855) gave the principles of least squares and the normal law of errors. Later on, in eighteenth, nineteenth and twentieth centuries, Euler, Lagrange, Bayes, Markov, Khintchine and Kolmogorov also developed the theory of probability.

Sir F. Galton (1822-1921) gave the concept of regression line. He and his successor Karl Pearson (1837-1936) are the pioneer of correlational analysis and  $\chi^2$ -test which play an important role in modern theory of statistics. W. S. Gosset discovered the Student's distribution which started a new era of exact small sample tests.

Sir R. A. Fisher (1890-1962) who is popularly known as the father of statistics, made a number of original work which gave a sound footing of the subject 'Statistics' in the diversified field such as Genetics, Biometry,

Education, Agriculture etc. He is the pioneer in introducing the concept of point estimation (efficiency, sufficiency, principle of maximum likelihood etc.) fiducial inference and exact sampling distributions. He along with F. Yates made a remarkable contribution in the field of 'Analysis of Variance' and 'Design of Experiments'. O. H. Hartley and P. C. Mahalanabis (1893-1972) contribute significant works in the field of sample survey. The above contributions placed Statistics in a very significant position among sciences.

## 1.2 Definitions of Statistics

Different authors defined statistics in a number of ways. Among those some of the important definitions are given below :

Webster defined statistics as "Statistics are classified facts representing the condition of the people in a state specially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement."

Dr. A. L. Bowley defined "Statistics are numerical statement of facts in any department of enquiry placed in relation to each other".

According to Yule and Kendall, "By statistics, we mean quantitative data affected to a marked extent by a multiplicity of causes."

A more exhaustive definitions of statistics is given by Prof. H. Secrist as "By statistics we mean aggregate of facts affected to a marked extent by a multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other."

## 1.3 Uses of Statistics in Different Fields

Now-a-days statistics is not only used for collecting numerical data but also to develop sound techniques for their handling, analysis and drawing valid inference from them. It is now used widely in different spheres of life—social, political and also in different fields such as Agriculture, Planning, Biology, Psychology, Education, Economics, Business Management etc. In short, following are the different important fields where statistics can be extensively used.

i) **Agriculture** : To have information in regards to total production of a certain agricultural product, total cultivable land, consumption of different types of crops, different types and levels of irrigations for different types of agricultural products, different doses of fertilizers and distribution of the



## Introduction

optimum doses of different fertilizers, livestock resources and their development etc. are usually obtained by agricultural census which mainly follow statistical methodology. With the progress of time, different new varieties are developed. To choose a new variety which mainly suits our climatic conditions, with the available resource we can get maximum production and to determine the optimum doses of fertilizers and measurement of irrigated water levels, there is a branch of statistics called Design of Experiments. This can be applied to different fields such as Dairy, Poultry, Chemical Industry etc.

For a meaningful and correct decision regarding land reforms, a government should know the actual distribution of land holdings. In short, 'Agricultural Statistics' may play a key role in agricultural development.

ii) **Planning** : At this age of planning, a government has to take help of the subject Statistics in case of doing any fruitful planning and policy formulation for building sound economy and economical development. Statistical tools are applied in knowing the rate of unemployment, cost of living expenditure, net profit in a certain management, for determining poverty line, rate of literacy etc. Thus in every successful planning, sound and correct analysis of complex statistical data are required.

iii) **Economics** : Number of economic problems such as wages, prices, analysis of time series data, demand analysis, cost and benefit analysis etc. require statistical data and proper use of statistical techniques. It also facilitates the development of the economic theory. Wide application of statistics and mathematics in the theory of economics led to new theories called, Economic Statistics and Econometrics.

iv) **Business** : In the domain of risk and uncertainties a successful businessman has to make proper use of statistics for making business-decisions. With the help of past records provided by statistics, a businessman should correctly estimate the demand so that the requirement of total raw material for a certain business period can be determined without uncertainty. For an example, an ice-cream manufacturer should have knowledge of the seasonal fluctuation of demand of his products and to assess the number of consumers by having an account of teenagers mainly.

v) **Industry** : In industry, statistics is widely used to provide quality control. In a production system, the quality of the product should be checked frequently so that the specification of the product is maintained. For this inspection plan, control chart etc. are extremely used. In inspection plan, a special type of sampling is adopted.

## An Introduction to the Theory of Statistics

iv) **Biology** : Biological research specially genetics and plant-breeding have wide application of statistical methods. R. A. Fisher made extensive use of statistics in biological research. He also developed a number of new statistical methods for analysing and interpreting data for generalization e.g. laws of variation heredity etc. A new branch of statistics named Biometry which mainly deals with the biological aspects may be mentioned here. Statistics helps demographic studies which includes the rates of birth and death, number of inhabitants, emigrants, immigrants, composition of families, construction of life tables etc.

vii) **Psychology & Education** : Statistics is widely used in education and psychology too, e.g. to determine the reliability and validity of a test, factor analysis etc. so much so that a new subject called Psychometry has come into existence in recent years.

viii) **Medicine** : Statistics is also used for the collection, presentation and analysis of observed facts relating to the causes of incidence of disease and the result obtained from the use of various drugs and medicines. The efficiency of manufactured drug or medicine is tested with the available drugs or medicines by using statistical methods.



## 2. VARIABLES AND FREQUENCY DISTRIBUTION

### 2.1 Population and Sample

Population means an aggregate of elements possessing certain characteristics of interest in any particular investigation or enquiry. It is generally named after the characteristics studied. Population may be finite or infinite. If we are interested to know the yield of certain crop of the individual farmer of Bangladesh, the aggregate of all relevant yields of the crop will constitute the population. Since all the elements are countable this type population is called finite population. Whereas all possible outcomes (Head and Tail) in successive tosses of a coin is an infinite population. It is evident from the above discussion that the statistical population differs from human population.

A sample is a representative part of the population. We are generally interested to know the properties of the population. Sometimes it is impractical or even impossible to handle the population because of limited resources. That is why, inferences about the population are usually drawn on the basis of the sample.

### 2.2 Variable

The measurements of elements of a population having certain characteristic may vary from element to element either in magnitude or in quality. These measurable characteristics are called variables.

There are two types of variables—qualitative and quantitative : Qualitative variables can be categorised in such a way that the categories must be mutually exclusive, whereas the quantitative variable can be measured. For example, the outcomes of a coin tossing problem gives qualitative variable with two categories—head or tail, the sex of persons had two categories—male or female. The categories are sometimes called attributes. The yield of crops, height of persons, the number of children in a family etc. may be considered as quantitative variables.

Again quantitative variables may be classified into two types namely, discrete variable and continuous variable. When the variable can assume only integral values, is called discrete variable. For example, the number of

children in a family. A variable is said to be continuous if it assumes any value within certain range. For example, the height of a person.

### 2.3 Frequency Distribution

Let us consider the marks out of 100 in Statistics obtained by 100 students in a certain examination of a University.

Table -2.1

Marks of Statistics of 100 Students of a Certain University.

54,	32,	38,	44,	48,	41,	30,	43,	46,	41,
47,	32,	26,	25,	41,	33,	51,	43,	45,	32,
51,	50,	34,	38,	44,	38,	54,	32,	39,	41,
42,	38,	41,	25,	45,	36,	40,	50,	52,	30,
41,	32,	27,	30,	40,	42,	52,	48,	49,	37,
48,	39,	26,	54,	47,	49,	38,	26,	27,	49,
47,	49,	32,	51,	49,	33,	47,	55,	25,	28,
37,	36,	44,	53,	48,	54,	29,	37,	39,	40,
50,	30,	55,	48,	36,	34,	27,	53,	28,	52,
47,	35,	46,	48,	32,	29,	54,	49,	47,	53.

The representation of the data in Table-2.1 do not provide us any useful information and may confuse us because of its large size. To condense these mass of data we used to prepare a table usually called frequency distribution which describes the pattern of the observations throughout its range.

Let us consider the marks as variable  $x$ . The above data are called raw data or ungrouped data. The condensation of the data without losing any information of interest is given in Table-2.2

Let us arrange the data in ascending or descending order of magnitude which is commonly termed as array. But this does not reduce the bulk of the data. A better representation is given in Table -2.2

A notation (/) which is usually called tally mark is put against each value of the variate  $x$ , when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (\) on the first four tallies. The technique facilitates the counting of the tally marks at the end. The total number of tally marks is known as frequency corresponding to the value of the variable.



## Variables and Frequency Distribution

**Table -2.2**  
**Frequency distribution of Marks of 100 Students.**

Marks	Tally Marks	Frequency	Marks	Tally Marks	Frequency
25	///	3	41	<del>///</del> ///	6
26	///	3	42	///	2
27	///	3	43	///	2
28	///	2	44	///	3
29	///	2	45	///	2
30	////	4	46	///	2
31		0	47	<del>///</del> ///	6
32	<del>///</del> ///	7	48	<del>///</del> ///	6
33	///	2	49	<del>///</del> ///	6
34	///	2	50	///	3
35	/	1	51	///	3
36	///	3	52	///	3
37	///	3	53	///	3
38	<del>///</del> ///	5	54	<del>///</del> ///	5
39	///	3	55	///	2
40	///	3			

In the Table -2.2 the frequency of the mark 38 is 5 i.e. 5 students got 38 marks. This representations, though better than any array does not condense the data to a great extent. Instead of considering the frequencies for each value of  $x$ , we can obtain the frequencies for a certain interval of marks, say,  $25 \leq x \leq 29$ ,  $30 \leq x \leq 34$  and so on. These intervals of the variable  $x$  are known as the class intervals of  $x$ . The lowest value of a class is called lower limit and the highest value of the same class is called the upper limit of the class interval. The difference between the upper and lower limits of the class is called the length of the class interval. The average of the lower and upper limits of a class interval is called the mid value of the class interval. The lower limit, upper limit and the mid-value of the class interval

$25 \leq x \leq 29$  are 25, 29 and  $\frac{25+29}{2} = 27$  respectively.

When either the lower limit of the first class interval or the upper limit of the last class interval or both are not specified, it is called open class interval. Open class intervals are sometimes seen in the frequency distribution of ages.

The raw data in Table-2.1 may further be condensed in the form of Table-2.3

**Table -2.3**  
**Frequency distribution of marks of 100 Students**

Class Interval of Marks	Tally Marks	Frequency
25—29	///	13
30—34	///	15
35—39	///	15
40—44	///	16
45—49	///	22
50—54	///	17
55—59	///	2

The frequency distribution in Table-2.3 is usually called discrete frequency distribution.

If we deal with a continuous variable, it is not possible to arrange the data in the class intervals of the above type. Let us consider the distribution of ages. If our intervals are 25—29, 30—34, then the person in the ages between 29 and 30 years can not be taken into consideration in such a case. To avoid this difficulty we may form the continuous class intervals with corresponding hypothetical frequencies as given below :

**Table -2.4**  
**Frequency Distribution of Ages**

Class intervals (Age in years)	Frequency
25—30	2
30—35	5
35—40	9
40—45	10
45—50	6
50—55	3

**Construction of a Frequency Distribution :** Following are the steps for the construction of a frequency distribution.

- 1) Find out range by subtracting the lowest value from the highest value of the variable  $x$ .
- 2) The number of class intervals should not be too large or too small, usually it lies between 5 and 20, considering the practical situation. Having fixed the number of classes, divide the range by it and the nearest integer to this value gives the length of class interval. The class intervals should be exhaustive, mutually exclusive and usually of equal length.



## Variables and Frequency Distribution

3) The table will have three columns having names—class interval, tally marks and frequency. The first class interval will start with the smallest value and continue until the interval with the highest value of the given series of data is reached.

4) Give tick mark to each of the values of the original table of raw data and put tally mark against the appropriate class interval. Thus exhaust all the values one after another. In case of continuous frequency distribution, the variable,  $x$  should follow either lower limit  $\leq x <$  upper limit or

lower limit  $< x \leq$  upper limit. The former of the limits is usually considered.

5) Count the number of tally marks corresponding to each class interval and write the result in the respective frequency column. For example see Table -2.3.

### 2.4 Graphical Representation of Frequency Distribution

Graphical representation of a frequency distribution is more effective than tabular representation, being easily understandable even to a lay-man. Diagrams are essential to convey the statistical information to the general public. It also facilitates the comparison of two or more frequency distributions.

The following types of graphs are generally used to represent the frequency distribution :

- i) Dot frequency diagram.
- ii) Histogram.
- iii) Frequency polygon.
- iv) Cumulative frequency polygon.
- v) Cumulative frequency curve or ogive.

i) **Dot frequency diagram** : In this diagram, we represent variable along  $x$  - axis and a dot along  $y$ -axis represents an observation. The number of dots corresponding to a certain value of the variable is the frequency of that value. Dot frequency diagram of frequency distribution given in Table-2.2 is shown in Fig. 2.1

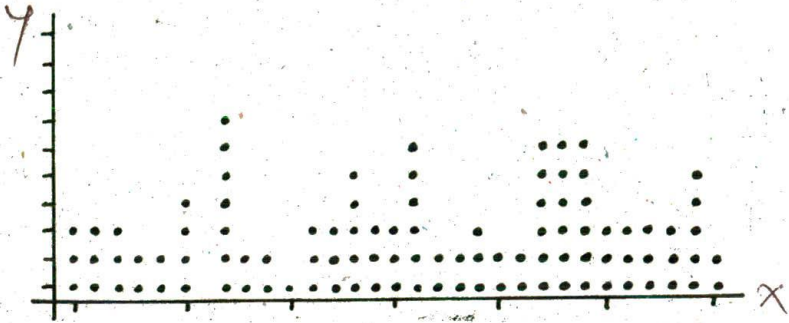


Fig. - 2.1 Dot frequency diagram .

ii) **Histogram** : In drawing histogram, the variable, expressed in continuous class intervals, are represented along x-axis and the frequencies along y - axis. On each class interval draw rectangle whose area is proportional to the frequency of the corresponding class interval. For equal class intervals the height of the rectangle is proportional to the frequency of the corresponding class interval. For unequal class interval, the height is proportional to the ratio of the frequency to the length of the class. The set of adjacent rectangles so constructed constitute the histogram. To draw histogram from an ungrouped distribution we have to assume the interval  $\left(x - \frac{h}{2}\right)$  to  $\left(x + \frac{h}{2}\right)$  where h is the jump from one value to the next. Similar modification can be carried out in classes of discrete frequency distribution. In that case, the class interval, becomes  $\left(l - \frac{h}{2}\right)$  to  $\left(u + \frac{h}{2}\right)$  where l and u indicate lower and upper limit of the class and h is the jump from one class to the next. The discrete frequency distribution given in Table- 2.3 can be arranged in continuous frequency distribution as in Table- 2.5.

Table-2.5  
Continuous frequency distribution of the marks of 100 students

Class Interval of marks	Mid values	Frequency	Cumulative frequency.
24.5—29.5	27	13	13
29.5—34.5	32	15	28
34.5—39.5	37	15	43
39.5—44.5	42	16	59
44.5—49.5	47	22	81
49.5—54.5	52	17	98
54.5—59.5	57	2	100



## Variables and Frequency Distribution

The histogram corresponding to Table - 2.5 is shown in Fig. 2.2

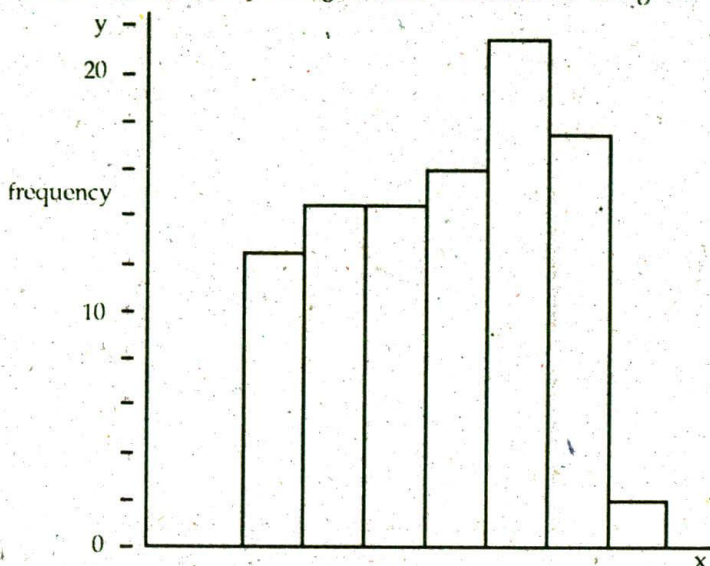


Fig. 2.2 Histogram

Histogram is a very popular graphical representation of the frequency distribution and is widely used.

A graph which is almost similar looking like a histogram is known as bar-diagram which may confuse a beginner. In bar diagram, different time periods or categories are represented along x - axis and their corresponding values are represented along y - axis. A bar diagram differs from a histogram in the following points :

- Histogram is used for continuous frequency distribution whereas bar-diagram is never used for that.
- In histogram the area of a rectangle is proportional to the frequency whereas in a bar-diagram the height of the bar is proportional to the value of the corresponding time period or category.
- The rectangles of a histogram are adjacent whereas the rectangles of a bar-diagram may or may not be adjacent.

✓ **iii) Frequency polygon :** In frequency polygon the mid-values of the continuous class intervals are represented along x-axis and the frequencies corresponding to the class intervals are represented along the y-axis. The class frequencies are plotted against the mid-values of the respective class

intervals. These points are then joint by straight lines one after another. The first and the last points are then brought down at each end to the x-axis by joining it to the mid-value of the next out lying interval of zero frequency. The polygon thus obtained is called frequency polygon. Frequency polygon of Table-2.5 is as given in Fig. 2.3.

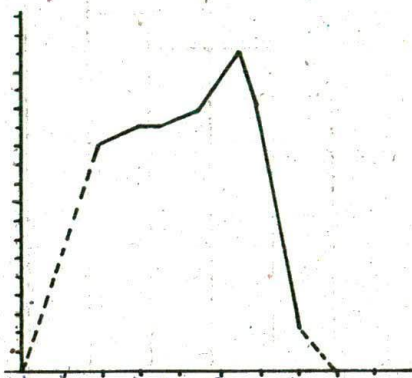


Fig. 2.3 Frequency polygon

ix) **Cumulative frequency polygon** : In cumulative frequency polygon the upper limits of the continuous class intervals are represented in x-axis and the cumulative frequencies are represented to the y-axis. The cumulative frequency means the cumulative total of the frequencies starting from the lowest class.

A cumulative frequency polygon is always non-decreasing but may be parallel to x-axis.

v) **Ogive** : A free hand curve to smooth a cumulative frequency polygon is called an ogive.

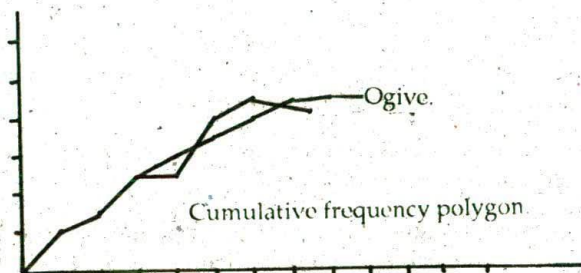


Fig. 2.4 Cumulative frequency polygon and ogive.



## 2.5 Frequency Curves

If the number of observations are large and the length of class intervals can be reduced, the frequency polygon will provide a smooth curve usually called frequency curve.

Following are the four different types of frequency curves.

- i) Symmetrical curve.
- ii) Moderately asymmetrical or skew curve.
- iii) Extremely asymmetrical or J-shaped curve.
- iv) U-shaped curve.

**i) Symmetrical curve:** A frequency curve is said to be symmetrical if the frequency at the mid-position is maximum and the rate of decrease from the peak of the curve is same in both the sides. Evidently it follows that if it can be folded along a vertical line, the two halves will coincide.

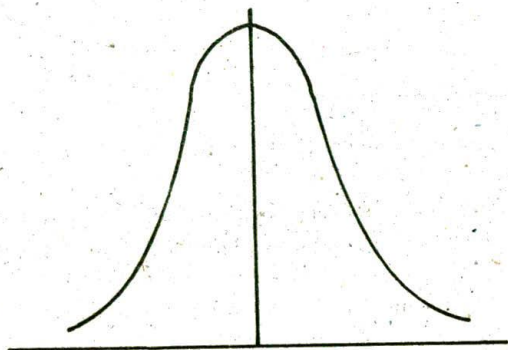


Fig. 2.5 Symmetrical curve.

**ii) Moderately asymmetrical or skew curve:** A frequency curve is said to be skew if it lacks in symmetry i.e. the rate of decrease from the peak point of the curve in both the sides are not equal. If the rate of decrease is rapid on the left side giving a longer tail at the right, we get a positively skew curve. For reverse case, the curve is said to be negatively skew.

Distribution of age of marriage follows a positively skew type of curve.



Fig. 2.6 (a) Positively skew curve. Fig. 2.6 (b) Negatively skew curve.

iii) **Extremely asymmetrical or J-shaped curve:** A frequency curve is said to be J-shaped if the maximum frequency occurs at one end of the distribution. If the maximum frequency occurs at the left end, then it gives a positively J-shaped curve. The distribution of wealth is usually represented by positively J-shaped curve.

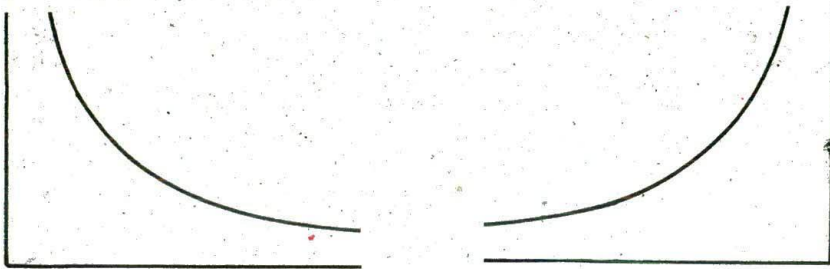


Fig. 2.7 (a) Positively J-shaped curve.

Fig. 2.7 (b) Negatively J-shaped curve.

iv) **U-shaped curve:** A frequency curve is said to be U-shaped if it looks like the letter U. In this type of curve, the maximum frequency occurs at both end of the distribution while the minimum at the middle.

The distribution of human death follows a U-shaped curve.

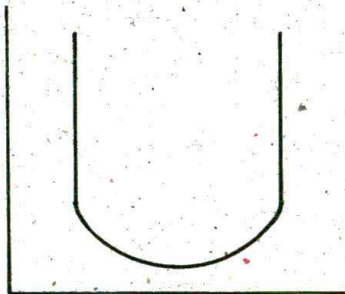


Fig. 2.8 U-shaped curve.

## 3. MEASURES OF LOCATION

### 3.1 Introduction

In a representative sample, the value of a series of data have a tendency to cluster around a certain point usually at the centre of the series. This tendency of clustering the values around the centre of the series is usually called central tendency. And its numerical measures are called the measures of central location.

### 3.2 Characteristics of an Ideal Measure of Location

- 1) It should be rigidly defined.
- 2) It should be readily comprehensible and easy to calculate.
- 3) It should be based upon all the observations.
- 4) It should be suitable for further algebraic treatments.
- 5) It should be affected as small as possible by sampling fluctuation.

### 3.3 Different Measures of Central Location

There are five different measures of central location:

- i) Arithmetic mean or Mean.
- ii) Geometric mean.
- iii) Harmonic mean.
- iv) Median.
- v) Mode.

**Arithmetic mean (AM):** Arithmetic mean of a set of observations is the sum of all observations divided by the number of observations e.g. the arithmetic mean or mean  $\bar{x}$  of  $n$  ungrouped observations  $x_1, x_2, \dots, x_n$  is given

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad \dots \dots (3.1)$$

**Example 3.1** Find arithmetic mean of 2, 5, 7, 9, 4 and 3

**Solution:** The arithmetic mean,  $\bar{x} = \frac{2 + 5 + 7 + 9 + 4 + 3}{6} = \frac{30}{6} = 5$ .

In case of frequency distribution (grouped data) as given in Table -3.1,



Table-3.1

Observation	frequency
$x_1$	$f_1$
$x_2$	$f_2$
$\vdots$	$\vdots$
$x_k$	$f_k$

the arithmetic mean,  $\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_kx_k}{f_1 + f_2 + \dots + f_k} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} = \frac{\sum_{i=1}^k f_i x_i}{n}$ , ..... (3.2)

where  $n = \sum_{i=1}^k f_i$

In case of grouped or continuous frequency distribution  $x_i$ 's are taken to be the mid-values of the intervals.

The method of calculation of mean in (3.2) is known as direct method.

**Short-cut method:** In this method, we can show the effect of change of scale and origin of the actual data. If  $x_i$  and  $f_i$  are large the calculation of  $\bar{x}$  by the formula given in (3.2) is time consuming and tedious. The calculation can be simplified by taking the deviations of the given values from any arbitrary value  $A$ , origin and dividing by  $h$ , scale which is generally the length of class interval in case of grouped frequency distribution. We define a new variate,  $u_i = \frac{x_i - A}{h}$

or,  $x_i = hu_i + A$  ..... (3.3)

with the help of (3.2) we get

$\bar{x} = \frac{\sum_{i=1}^k f_i (u_i + A)}{\sum_{i=1}^k f_i} = h \bar{u} + A$  .....(3.4)

which shows that arithmetic mean is dependent on change of origin and scale.

### Measures of Location

**Example 3.2** Find arithmetic mean by direct as well as short-cut method from the frequency distribution of wages with class interval of two Taka each from the following data of daily wages received by 35 labourers in a certain factory.

Class Interval of wages (Taka)	Number of labourers $f_i$
11—13	3
13—15	4
15—17	5
17—19	10
19—21	6
21—23	4
23—25	3
Total	35

**Solution :** Calculation of arithmetic mean by both the methods.

**Table-3.2**

Class Interval of wages (Taka)	Number of labourers $f_i$	Mid value of class Int. $x_i$	$f_i x_i$	New variate $u_i = \frac{x_i - 18}{2}$	$f_i u_i$
11—13	3	12	36	-3	-9
13—15	4	14	56	-2	-8
15—17	5	16	80	-1	-5
17—19	10	18	180	0	0
19—21	6	20	120	1	6
21—23	4	22	88	2	8
23—25	3	24	72	3	9
Total	35		632		1

**Direct method :**

$$\text{Arithmetic mean, } \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{632}{35} = 18.06 \text{ TK. (app.)}$$

**Short-cut method:**

Arithmetic mean  $\bar{x} = h \bar{u} + A$ , where  $h = 2$  and  $A = 18$

Now we calculate,  $\bar{u} = \frac{1}{35} = 0.03$ .  $[\bar{u} = \frac{\sum f_i u_i}{\sum f_i}]$

Therefore,  $\bar{x} = 0.03 \times 2 + 18 = 18.06$  TK. (app)

Hence the mean daily wages = 18.06 TK. (app) is obtained from both the methods.

**Property 1:** The sum of the deviations of set of observations from their

arithmetic mean is zero, i. e.  $\sum_{i=1}^k f_i (x_i - \bar{x}) = 0$  অর্থ  $\sum f_i x_i$  এর গুণিতক  $\sum f_i$  value  $\rightarrow n$

**Proof:** Let  $\bar{x}$  be the mean of  $x_i$  with frequencies  $f_i$  then

$$\sum_{i=1}^k f_i (x_i - \bar{x}) = \sum_{i=1}^k f_i x_i - \bar{x} \sum_{i=1}^k f_i = n \bar{x} - n \bar{x} = 0, \text{ where } n = \sum_{i=1}^k f_i$$

**Property 2:** The arithmetic mean of a set of  $N$  constant observations  $A$  is  $A$ .

**Proof:** Let us consider  $N$  constant observations which are  $A$ , say, then  $\bar{x} = \frac{\sum A_i}{N}$

$$\bar{x} = \frac{\sum A}{N} = \frac{NA}{N} = A.$$

$$\bar{x} = \frac{\sum A_i}{N} \Rightarrow n \bar{x} = \sum A_i$$

**Property 3:** The sum of the squares of the deviations of a set of observations is minimum when the deviations are taken about the arithmetic mean i. e.

we are to show,  $\sum_{i=1}^k f_i (x_i - A)^2 > \sum_{i=1}^k f_i (x_i - \bar{x})^2$  where  $A$  is any arbitrary constant.

**Proof:** Let  $\bar{x}$  be the arithmetic mean of a set of observations  $x_i$  with frequencies  $f_i$ , also let  $A$  be an arbitrary value, we have,

$$\begin{aligned} \sum_{i=1}^k f_i (x_i - A)^2 &= \sum_{i=1}^k f_i (x_i - \bar{x} + \bar{x} - A)^2 \\ &= \sum_{i=1}^k f_i (x_i - \bar{x})^2 + n (\bar{x} - A)^2 + 2 (\bar{x} - A) \sum_{i=1}^k f_i (x_i - \bar{x}) \end{aligned}$$

$$\sum_{i=1}^k f_i (x_i - \bar{x}) = 0$$

third term vanishes due to property 1. Hence we get,

$$\sum_{i=1}^k f_i (x_i - A)^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2 + n (\bar{x} - A)^2. \text{ As } n (\bar{x} - A)^2 \text{ is a positive quantity}$$

$$\sum_{i=1}^k f_i (x_i - A)^2 > \sum_{i=1}^k f_i (x_i - \bar{x})^2. \text{ Hence proved.}$$



Ass

### Measures of Location

**Property 4:** (Mean of the composite series). If  $\bar{x}_i$  ( $i = 1, 2, \dots, k$ ) are the means of  $k$  series of sizes  $n_i$  ( $i = 1, 2, \dots, k$ ) respectively, then the mean  $\bar{x}$  of the composite series obtained by the formula.

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots + n_k}$$

**Proof:** Let  $x_{11}, x_{12}, \dots, x_{1n_1}$  be  $n_1$  numbers in the first series,  $x_{21}, x_{22}, \dots, x_{2n_2}$  be  $n_2$  numbers in the second series and so on  $x_{k1}, x_{k2}, \dots, x_{kn_k}$  be  $n_k$  numbers in the  $k$ th series. By the formula given in (3.1) we have,

$$\bar{x}_1 = \frac{x_{11} + x_{12} + \dots + x_{1n_1}}{n_1}$$

$$\bar{x}_2 = \frac{x_{21} + x_{22} + \dots + x_{2n_2}}{n_2} \text{ and so on}$$

$$\bar{x}_k = \frac{x_{k1} + x_{k2} + \dots + x_{kn_k}}{n_k}$$

The arithmetic mean  $\bar{x}$  of the composite series of size  $n_1 + n_2 + \dots + n_k$  is give by

$$\bar{x} = \frac{(x_{11} + x_{12} + \dots + x_{1n_1}) + (x_{21} + x_{22} + \dots + x_{2n_2}) + \dots + (x_{k1} + x_{k2} + \dots + x_{kn_k})}{n_1 + n_2 + \dots + n_k}$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{\sum_{i=1}^k n_i}$$

#### Merits :

- (1) It is rigidly defined and easy to calculate.
- (2) It is easy to understand and easy for algebraic treatment.
- (3) It takes all the observations into account.
- (4) It is less affected by sampling fluctuation.

#### Demerits :

- (1) It is affected by extreme values.
- (2) It is impossible to calculate if the extreme classes of the frequency distribution are open.
- (3) The value of the arithmetic mean may not occur in the series.

(ii) **Geometric mean** : The geometric mean of a set of n non-zero positive observations is the nth root of their product. Let  $x_1, x_2, \dots, x_n$  be n non-zero positive observations in a series of data.

Thus the geometric mean,  $G = (x_1 \times x_2 \times \dots \times x_n)^{\frac{1}{n}}$  .....(3.5)

For example, the geometric mean, G of 2, 4 and 8 is.

$(2 \times 4 \times 8)^{\frac{1}{3}} = (64)^{\frac{1}{3}} = 4$ .

The calculation may sometime be simplified by taking logarithm , that is,

$\log G = \frac{1}{n} [\log x_1 + \log x_2 + \dots + \log x_n] = \frac{\sum \log x_i}{n}$  .....(3.6)

Thus log G is arithmetic mean of log  $x_i$  s. The antilog of log G will give the value of G. If the observations are given in frequency distribution (grouped data) as given in Table 3.1, then the geometric mean is given by,

$G = \left( x_1^{f_1} \times x_2^{f_2} \times \dots \times x_k^{f_k} \right)^{\frac{1}{n}}$  ; where  $n = \sum f_i$  .....(3.7)

Here also,  $\log G = \frac{1}{n} \sum_{i=1}^k f_i \log x_i$  .....(3.8)

Thus the value of G in a frequency distribution can also be obtained by considering  $x_i$ 's as the mid values of the class intervals.

**Example 3.3** Find geometric mean of the frequency distribution given in Example 3.2

**Solution :**

**Table-3.3**

Class Interval of wages (Taka)	Number of Labourers $f_i$	Mid values of class Int. $x_i$	$\log x_i$	$f_i \log x_i$
11—13	3	12	1.0792	3.2376
13—15	4	14	1.1461	4.5844
15—17	5	16	1.2041	6.0205
17—19	10	18	1.2553	12.5530
19—21	6	20	1.3010	7.8060
21—23	4	22	1.3424	5.3696
23—25	3	24	1.3802	4.1406
Total	35			43.7117

## Measures of Location

We know,  $\log G = \frac{\sum_{i=1}^k f_i \log x_i}{\sum f_i} = \frac{43.7117}{35} = 1.2489$

Therefore,  $G = 17.74$  Tk. (app)

*anti log*

*43.7117*

### Merits :

- 1) It is rigidly defined.
- 2) It takes all the observations into account.
- 3) It is not affected much by sampling fluctuation.
- 4) It gives comparatively more weights to small observations.

### Demerits :

- 1) It is difficult to understand and to calculate for a student with less mathematical knowledge.
- 2) It is impossible to calculate if the extreme classes of the frequency distribution are open.
- 3) The value of the geometric mean may not occur in the series.

Uses : Geometric mean is mainly used—

- 1) to calculate averages of ratios and percentages ;
- 2) for the construction of index numbers.

(iii) **Harmonic mean** : The harmonic mean of a set of  $n$  non-zero observations  $x_1, x_2, \dots, x_n$  in a series is the reciprocal of the arithmetic mean of the reciprocals.

Thus the harmonic mean,  $H = \frac{1}{\frac{1}{n} \sum \frac{1}{x_i}}$  ..... (3.9)

For example, the harmonic mean of 4, 5 and 9 is

$$H = \frac{1}{\frac{1}{3} \left( \frac{1}{4} + \frac{1}{5} + \frac{1}{9} \right)} = \frac{1}{\frac{1}{3} (.2500 + .2000 + .1111)} = \frac{1}{.1870} = 5.35 \text{ (app)}$$

In case of frequency distribution given in Table 3.1.

$$H = \frac{1}{\frac{1}{n} \sum \frac{f_i}{x_i}} \text{ where } n = \sum_{i=1}^k f_i \text{ ..... (3.10)}$$

$x_i$ 's may be considered as the mid-values of the class intervals.



**Example 3.4** Find the harmonic mean of the frequency distribution given in Example 3.2.

**Solution :**

**Table-3.4**

Class Interval of wages (Taka)	Number of Labourers ( $f_i$ )	Mid values of class Int. ( $x_i$ )	$\frac{1}{x_i}$	$\frac{f_i}{x_i}$
11—13	3	12	.0833	.2499
13—15	4	14	.0714	.2856
15—17	5	16	.0625	.3125
17—19	10	18	.0556	.5560
19—21	6	20	.0500	.3000
21—23	4	22	.0455	.2176
23—25	3	24	.0417	.1251
Total	35			2.0467

The harmonic mean  $\frac{1}{\frac{2.0467}{35}} = \frac{35}{2.0467} = 17.10$  Tk. (app)

**Merits :**

- 1) It is rigidly defined.
- 2) It takes all the observations into account.
- 3) It is not affected much by small observations.

**Demerits :**

- 1) It is not easy to understand and difficult to calculate.
- 2) It is impossible to calculate if the extreme classes of the frequency distribution are open.
- 3) The value of the harmonic mean may not occur in the series.

**Uses :** The harmonic mean is used when the observation are expressed in terms of rates, speeds, prices etc.

**Relationship Between Arithmetic Mean, Geometric Mean and Harmonic Mean.**

**Theorem 3.1** For two non-zero positive observations  $AH=C^2$  where  $A$  = Arithmetic mean,  $H$  = Harmonic mean, and  $C$  = Geometric mean.

**Proof.** Let the two observations be  $x_1$  and  $x_2$

$$\text{then } A = \frac{x_1 + x_2}{2}, C = (x_1 x_2)^{\frac{1}{2}} \text{ and } H = \frac{1}{\frac{1}{2} \left( \frac{1}{x_1} + \frac{1}{x_2} \right)} = \frac{2x_1 x_2}{x_1 + x_2}$$

### Measures Of Location

Therefore,  $AH = \frac{(x_1 + x_2)}{2} \times \frac{2x_1x_2}{(x_1 + x_2)} = x_1x_2 = G^2$ . Hence proved.

**Theorem 3.2** For n non-zero positive observations,

Arithmetic mean  $\geq$  Geometric mean  $\geq$  Harmonic mean.

**Proof :** Let  $x_1, x_2, \dots, x_n$  be n non-zero positive observations. Also let A, H and G are as defined in Theorem 3.1 and  $d_i = x_i - A$ .

we know,  $G = (x_1 \times x_2 \times \dots \times x_n)^{\frac{1}{n}}$

Taking logarithm on both sides, we have  $\log G = \frac{1}{n} \sum_{i=1}^n \log x_i$

$$= \frac{1}{n} \sum_{i=1}^n \log (A + d_i) = \frac{1}{n} \sum_{i=1}^n \log A \left( 1 + \frac{d_i}{A} \right)$$

$$= \log A + \frac{1}{n} \sum_{i=1}^n \log \left( 1 + \frac{d_i}{A} \right)$$

Expanding  $\log \left( 1 + \frac{d_i}{A} \right)$  in ascending power of  $\frac{d_i}{A}$  by Taylor's expansion method and avoiding 3rd or more power we have,

$$\log \left( 1 + \frac{d_i}{A} \right) = \frac{d_i}{A} - \frac{\left( \frac{d_i}{A} \right)^2}{2 \left( 1 + \theta \frac{d_i}{A} \right)^2} ; 0 \leq \theta \leq 1$$

$$\text{Therefore, we get, } \log G = \log A + \frac{1}{n} \sum_{i=1}^n \frac{d_i}{A} - \frac{1}{n} \sum_{i=1}^n \frac{\left( \frac{d_i}{A} \right)^2}{2 \left( 1 + \theta \frac{d_i}{A} \right)^2} ; 0 \leq \theta \leq 1$$

=  $\log A + 0$  - a positive quantity

$\therefore \log G \leq \log A$  or  $A \geq G$

..... (3.11)

Again we know

$$\frac{1}{H} = \frac{1}{n} \left( \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right) \geq \left( \frac{1}{x_1} \times \frac{1}{x_2} \times \dots \times \frac{1}{x_n} \right)^{\frac{1}{n}} \geq \frac{1}{G}$$

$$\therefore G \geq H \quad \dots \dots (3.12)$$

combining (3.11) and (3.12) we have,  $A \geq G \geq H$ . Hence proved.

The equality holds when all  $x$ 's are equal. For example, for a set of observations say, 5, 5, 5, 5, the arithmetic mean, the geometric mean and the harmonic mean give the same value equal to 5.

**(ii) Median:** The median is defined as the middle most observation when the observations are arranged in order of magnitude.

For ungrouped data, when  $n$  is odd, the middle most observation i. e. the

$\left( \frac{n+1}{2} \right)$  th observation will be the median in the series.

*Handwritten notes:*  
 (3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)  
 (21, 22, 23, 24, 25, 26, 27, 28, 29, 30)  
 (31, 32, 33, 34, 35, 36, 37, 38, 39, 40)  
 (41, 42, 43, 44, 45, 46, 47, 48, 49, 50)

Again when  $n$  is even, the median will be the arithmetic mean of  $\frac{n}{2}$  th and

$\left( \frac{n}{2} + 1 \right)$  th observations in the series.

For example, the median of the observations 5, 10, 7, 3, 2, i. e. 10, 7, 5, 3, 2, is 5 and the median of 11, 3, 9, 5, 7, 13, i. e. 13, 11, 9, 7, 5, 3, is

$$\frac{9+7}{2} = 8. \text{ For grouped frequency distribution the median is given by}$$

$$Me = L + \frac{\frac{n}{2} - F}{f} \times c;$$

where,  $L$  = the lower limit of the median class (median class is that class which contains  $\frac{n}{2}$  th observation of the series);

$N$  = total number of observation;

$F$  = cumulative frequency of the class just preceding the median class,

$f$  = frequency of the median class, and

$c$  = length of the median class.



## Measures Of Location

**Example 3.5** Find median from the frequency distribution given in Example 3.2.

**Solution**

**Table-3.5**

Class Interval of wages (Taka)	Number of labours $f_i$ (frequency)	Cumulative frequency
11—13	3	3
13—15	4	7
15—17	5	12
17—19	10	22
19—21	6	28
21—23	4	32
23—25	3	35

*Handwritten notes in the table:*  
 -  $\Delta_1$  diff: arrow from 4 to 5  
 -  $\Delta_2$  diff: arrow from 5 to 10  
 -  $t_c$ : arrow from 7 to 12  
 -  $F$ : arrow from 12 to 22  
 -  $n = 35$  written on the right side of the table.

Here  $n = 35$ , (17—19) is the median class i.e.  $\frac{35}{2}$  th is 17.5th observation lies in that class.

Therefore,  $Mc = 17 + \frac{17.5 - 12}{10} \times 2 = 17 + 1.1 = 18.1$  Tk.

**Merits :**

- 1) Median is rigidly defined
- 2) It is easily understood and easy to calculate.
- 3) It is not all affected by extreme values.
- 4) It can be calculated from frequency distribution with open end.

**Demerits :**

- 1) In case of even number of observations, median cannot be defined exactly.
- 2) It is not based on all the observations.
- 3) It is not easy for algebraic treatment.
- 4) It is affected much by sampling fluctuation.

**Mode :** The mode is that observation of the variable for which the frequency is maximum.

For example, the mode of the observations 2, 5, 9, 5, 3, 5 is 5.

For grouped frequency distribution the mode is given by,

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c;$$

where,  $L$  = the lower limit of the modal class (modal class is that class for which the frequency is maximum).

$\Delta_1$  = the differences between the frequency of the modal class and pre-modal class.

$\Delta_2$  = the differences between the frequency of the modal class and post-modal class,

and  $c$  = the length of the modal class.

**Example 3.6** Find mode from the frequency distribution given in Example 3.2.

**Solution :** Here the modal class is (17—19) because in that class the frequency is maximum i. e. 10

Therefore,  $M_o = 17 + \frac{5}{5+4} \times 2 = 17 + 1.11 = 18.11$  Tk. (app)

**Merits :**

- 1) Mode is easy to understand and easy to calculate.
- 2) It is not at all affected by extreme values.
- 3) It can be calculated from frequency distribution with open class.

**Demerits :**

- 1) Mode is not clearly defined in case of bi-modal or multimodal distribution.
- 2) It is not based on all the observations.
- 3) It is difficult for algebraic treatments.
- 4) It is affected to a great extent by sampling fluctuation.

### 3.4 Other Measures of Location : Quartiles, Deciles and Percentiles

Quantiles are those values in a series which divide the total frequency into number of equal parts when the series is arranged in order of magnitude. Some important quantiles are quartiles, deciles and percentiles.

Quartiles are those values which divide the total frequency into four equal parts. The value of the quartile having the position mid-way between the lower extreme and the median is the first quartile and is denoted by  $Q_1$  and that between the median and the upper extreme is the third quartile and is denoted by  $Q_3$ . The median is thus one of the quartiles and is denoted by  $Q_2$ .

For a grouped frequency distribution the quartiles are given by

## Measures Of Location

$$Q_i = L_i + \frac{\frac{i \times n}{4} - F_i}{f_i} \times c; \quad i=1, 2, 3.$$

where,  $L_i$  = lower limit of the  $i$ th quartile class ( $i$ th quartile class is that class which contains the  $\frac{i \times n}{4}$ th observation)

$n$  = total number of observation

$F_i$  = Cumulative frequency of the pre- $i$ th quartile class

$f_i$  = frequency of the  $i$ th quartile class, and

$c$  = length of class interval of the  $i$ th quartile class.

Deciles and percentiles are those values which divide the total frequency into 10 and 100 equal parts respectively. The median is the 5th decile,  $D_5$  and 50th percentile,  $P_{50}$ .

For grouped frequency distribution, the deciles are given by

$$D_j = L_j + \frac{\frac{j \times n}{10} - F_j}{f_j} \times c; \quad j=1, 2, 3, \dots, 9,$$

where,  $L_j$  = lower limit of the  $j$ th decile class, ( $j$ th decile class contains the  $\frac{j \times n}{10}$ th observation);

$n$  = total number of observation;

$F_j$  = cumulative frequency of the pre- $j$ th decile class;

$f_j$  = frequency of the  $j$ th decile class;

and  $c$  = length of class interval of the  $j$ th decile class.

For grouped frequency distribution the percentiles are given by

$$P_k = L_k + \frac{\frac{k \times n}{100} - F_k}{f_k} \times c; \quad k = 1, 2, 3, \dots, 99,$$

where,  $L_k$  = lower limit of the  $k$ th percentile class ( $k$ th percentile class is that class which contains the  $\frac{(k \times n)}{100}$ th observation);

$n$  = total number of observation;

$F_k$  = cumulative frequency of the pre- $k$ th percentile class;

$f_k$  = frequency of the  $k$ th percentile class, and

$c$  = length of class interval of the  $k$ th percentile class.

**Example 3.7** Find a) first and third quartile b) 7th decile and c) 62th percentile from the frequency distribution given in Example 3.2

**Solution :** a) (15—17) is the first quartile ( $Q_1$ ) class because



$\frac{n}{4}$ th =  $\frac{35}{4}$ th = 8.75th observation lies in that class.

Hence,  $Q_1 = 15 + \frac{8.75 - 7}{5} \times 2 = 15 + 0.7 = 15.7$  Tk.

(19—21) is the third quartile ( $Q_3$ ) class because  $\frac{3n}{4}$ th = 26.25th observation lies in that class.

Hence,  $Q_3 = 19 + \frac{26.25 - 22}{6} \times 2 = 19 + 1.42 = 20.42$  Tk. (app)

b) (19—21) is the 7th decile ( $D_7$ ) class as  $\frac{7n}{10}$ th = 24.5 th observation lies in the class.

Hence,  $D_7 = 19 + \frac{24.5 - 22}{6} \times 2 = 19 + 0.83 = 19.83$  Tk. (app.)

c) (17—19) is the 62th percentile ( $P_{62}$ ) class as  $\frac{62 \times n}{100}$ th = 21.7th observation lies in that class.

Hence,  $P_{62} = 17 + \frac{21.7 - 12}{10} \times 2 = 18.94$  Tk. (app)

### 3.5 Graphical Determination of Mode and Quantiles

The value of the mode can be determined graphically from the histogram with equal length of class intervals since the value of mode lies within the tallest rectangle of the distribution. The method uses three adjacent rectangles of the histogram with the tallest in the middle. The mode is the abscissa of the point P at which AB and CD intersect. (See Fig. 3.1)

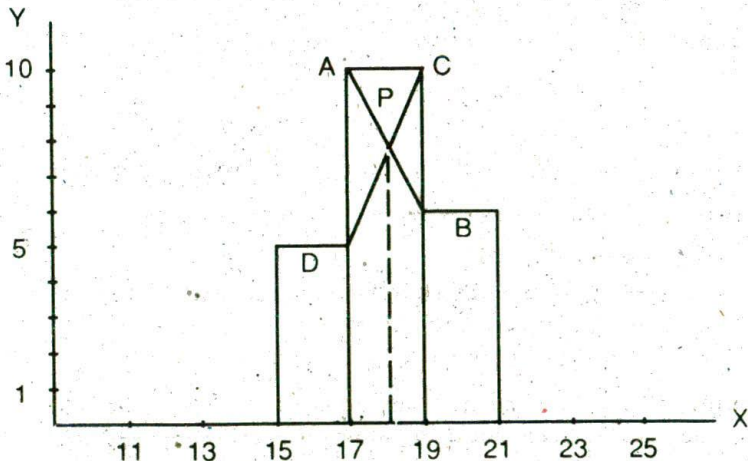


Fig. 3.1. Determination of mode graphically.

## Measures Of Location

For finding the values of median and other quantiles graphically, we are to draw a cumulative frequency polygon or ogive. To determine the value of the median we mark a point along the y-axis corresponding to  $\frac{n}{2}$ . From this point we draw a line parallel to the x-axis and mark the point where this line intersect the curve. Then a perpendicular is drawn from the point of intersection to the abscissa. The distance between the origin and the foot of the perpendicular gives the median.

For finding the value of  $Q_1$  and  $Q_3$ , we take points on the y-axis corresponding to  $\frac{n}{4}$  and  $\frac{3n}{4}$  respectively and proceed as above.

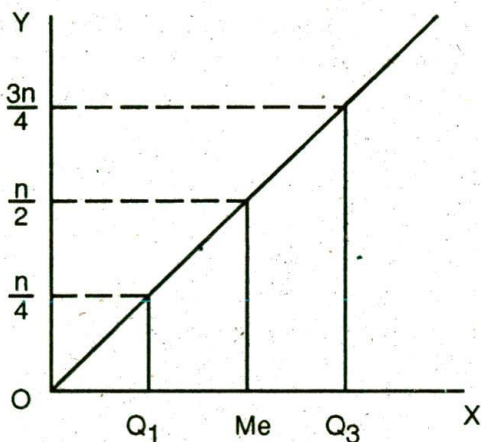


Fig. 3.2 Graphical determination of quartile

This method is applied for determining other quantile values also.

### 3.6 Empirical Relationship Among Arithmetic Mean, Median and Mode

For moderately asymmetrical distribution, the empirical relationship among the arithmetic mean, median and mode is,

$$\text{Arithmetic mean—Mode} = 3 (\text{Arithmetic mean—Median})$$

For a positively skew distribution the relative position of arithmetic mean, median and mode is shown graphically in Fig. 3.3.

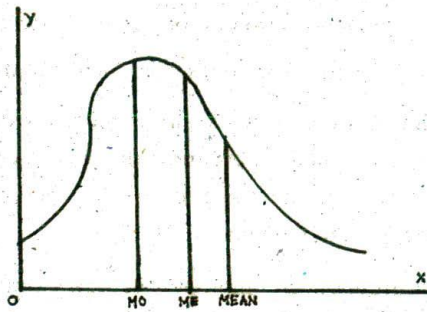


Fig. 3.3 Relative position of mean, median and Mode in a positively skew distribution.

For negatively skew distribution the relative position of mean, median and mode will be reverse. For a symmetrical distribution mean, median and mode coincide.



## 4. MEASURES OF DISPERSION

### 4.1 Introduction

Measures of central location give us an idea of the concentration of the observations about the central value of the distribution. It is equally important to know how the observations of the variate cluster around or dispersed away from the central value of the distribution. Let us consider two groups each of 6 students with their scores in a particular examination.

Gr—I	48,	50,	52,	51,	49,	50
Gr—II	1,	2,	100,	99,	98,	0

The arithmetic mean for each group is 50. It is very much apparent from the data that the first group consists of average or near average intelligent students and the second group is made up of very bright and very dull students. Graphically the above phenomenon can be shown as below :

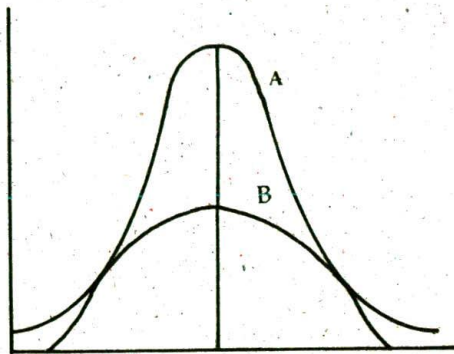


Fig-4.1 Comparison of dispersion of two distributions.

It is evident that the distribution A and B have the same arithmetic mean  $\bar{x}$ , but they differ in variation from  $\bar{x}$ . Such variation is usually called dispersion. Measures of dispersion give the degree of scatterness about the central location and thus giving measure of variability or lack of homogeneity of the data.

### 4.2 Characteristics of an Ideal Measure of Dispersion

Following are the characteristics of an ideal measure of dispersion.

- 1) It should be rigidly defined.

- 2) It should be easy to calculate and easy to understand.
- 3) It should be based on all the observations.
- 4) It should be amenable to further algebraic treatments.
- 5) It should be less affected by sampling fluctuation.

### 4.3 Measures of Dispersion

Following are the measures of dispersion :

#### a) Absolute measures

1. Range.
2. Quartile deviation.
3. Standard deviation.
4. Mean deviation.

#### b) Relative measures

1. Co-efficient of quartile deviation.
2. Co-efficient of variation.
3. Co-efficient of mean deviation.

### Absolute Measures of Dispersion

**1. Range:** Range is the difference of the highest and the lowest observations of the distribution.

The range is  $52-48=4$  for group 1 and  $100-0=100$  for group 2 of the students given above.

#### Merits :

- 1) It is the simplest measure of dispersion.
- 2) It is easy to calculate and easy to understand.
- 3) It is based on the extreme observations only and no detail information is required.
- 4) It gives us a quick idea of the variability of the observations involving least amount of time and calculations.

#### Demerits :

- 1) It is a crude measure of dispersion as the two extreme observations may be subject to sampling fluctuation.
- 2) It would be misleading if any of the two extreme values has very high or low magnitude.
- 3) It cannot be calculated if the extreme classes of the frequency distribution are open.

#### Uses of Range :

- 1) It is used in measurement of share market fluctuation.
- 2) It is also used in statistical quality control work.

## Measures of Dispersion

### 2. Quartile Deviation or Semi-interquartile Range

Quartiles divide the observations into 4 equal parts when the observations are arranged in order of magnitude. Median which may be denoted by  $Q_2$  is the middle most observation and  $Q_1$  and  $Q_3$  are the middle most observation of the lower half and the upper half respectively. Therefore,  $Q_2 - Q_1$  and  $Q_3 - Q_2$  give us some measures of dispersion. The arithmetic mean of these two measures gives us the quartile deviation or semi-interquartile range, denoted by  $Q$ .

$$\text{That is, } Q = \frac{(Q_2 - Q_1) + (Q_3 - Q_2)}{2} = \frac{Q_3 - Q_1}{2} \quad \dots\dots (4.1)$$

**Example 4.1** Find out quartile deviation from the following frequency distribution.

Variable	Frequency
0—5	2
5—10	5
10—15	7
15—20	13
20—25	21
25—30	16
30—35	8
35—40	3

**Solution :** We can obtain the quartile values easily as

$$Q_1 = 16.83, Q_2 = 22.5 \text{ and } Q_3 = 27.58$$

$$\text{Therefore, } Q = \frac{Q_3 - Q_1}{2} = \frac{10.75}{2} = 5.375$$

**Remark :** Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data it cannot be regarded as a reliable measure.

**3. Standard Deviation :** The arithmetic mean of the squares of the deviations of the given observations from their arithmetic mean is known as variance. The positive square root of variance is the standard deviation.



### An Introduction to The Theory of Statistics

If  $x_1, x_2, \dots, x_n$  be  $n$  observations of a variable, then standard deviation is

$$\text{defined by } s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \dots \dots (4.2)$$

In case of frequency distribution or grouped data

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^k f_i (x_i - \bar{x})^2}; \quad \dots \dots (4.3)$$

where  $n = \sum f_i$  and  $\bar{x}$  is the arithmetic mean of the distribution.

**Working formula :** The quantity  $\sum_{i=1}^n (x_i - \bar{x})^2$  is called the sum of squares of  $x$ 's.

$$\text{Now, } \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2 \bar{x} x_i + \bar{x}^2)$$

$$= \sum_{i=1}^n x_i^2 - 2 \bar{x} \sum_{i=1}^n x_i + n \bar{x}^2 = \sum_{i=1}^n x_i^2 - n \bar{x}^2$$

$$= \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \quad \dots \dots (4.4)$$

The term  $\sum_{i=1}^n x_i^2$  is called raw sum of squares and  $\frac{(\sum_{i=1}^n x_i)^2}{n}$  is called

correction factor.

Proceeding as above in case of grouped data we get,

$$\sum_{i=1}^k f_i (x_i - \bar{x})^2 = \sum_{i=1}^k f_i x_i^2 - \frac{\left(\sum_{i=1}^k f_i x_i\right)^2}{n}; \text{ where } \sum_{i=1}^k f_i = n. \quad \dots \dots (4.5)$$

## Measures of Dispersion

Thus the working formula for standard deviation is given by,

$$s = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}, \text{ for ungrouped data} \quad \dots \dots \dots (4.6)$$

and

$$s = \sqrt{\frac{\sum f_i x_i^2}{n} - \left(\frac{\sum f_i x_i}{n}\right)^2}; \text{ where } n = \sum f_i; \text{ for grouped data} \quad \dots \dots (4.7)$$

**Theorem 4.1** Standard deviation is independent on change of origin but not of scale.

**Proof :** Let  $x_1, x_2, \dots, \dots, x_k$  be the mid-values of the classes of a frequency distribution and let  $f_1, f_2, \dots, \dots, f_k$  be their corresponding frequencies.

$$\text{Also let, } u_i = \frac{x_i - A}{h}$$

where  $u_i, A$  and  $h$  are changed variate, origin and scale respectively.

Now,  $x_i = hu_i + A$  or,  $\bar{x} = h \bar{u} + A$

Putting the values of  $x_i$  and  $\bar{x}$  in the formula given in (4.3)

we have,

$$s_x = \sqrt{\frac{1}{n} \sum f_i (hu_i + A - h \bar{u} - A)^2} = \sqrt{\frac{h^2}{n} \sum f_i (u_i - \bar{u})^2}$$

$$= \sqrt{h^2 s_u^2} \text{ where } s_u^2 \text{ is the variance of } u \text{ variate.}$$

$$\therefore s_x = |h s_u| \quad \dots \dots \dots (4.8)$$

showing that standard deviation is independent on change of origin but not of scale.

**Note :** This method of calculation of standard deviation is known as short-cut method while the earlier method is known as direct method.

**Example 4.2** Calculate standard deviation from the frequency distribution given in Example 4.1 by a) direct method and b) short-cut method.

**Solution :** Let us prepare a table for calculation of standard deviation by both the methods.

**Table -4.1**

Class Interval	Mid values $x_i$	Frequency $f_i$	$f_i x_i$	$f_i x_i^2$	$u_i = \frac{x_i - 22.5}{5}$	$f_i u_i$	$f_i u_i^2$
0—5	2.5	2	5.0	12.50	-4	-8	32
5—10	7.5	5	37.5	281.25	-3	-15	45
10—15	12.5	7	87.5	1093.75	-2	-14	28
15—20	17.5	13	227.5	3981.25	-1	-13	13
20—25	22.5	21	472.5	10631.25	0	0	0
25—30	27.5	16	440.0	12100.00	1	16	16
30—35	32.5	8	260.0	8450.00	2	16	32
35—40	37.5	3	112.5	4218.75	3	9	27
<b>Total</b>		<b>75</b>	<b>1642.5</b>	<b>40768.75</b>		<b>-9</b>	<b>193</b>

a) **Direct method :** From (4.7) we have,

$$s_x = \sqrt{\frac{40768.75}{75} - \left(\frac{1642.5}{75}\right)^2} = \sqrt{543.58 - 479.61}$$

$$= \sqrt{63.97} = 7.99 \text{ (app).}$$

b) **Short-cut method :** From (4.8) we have,

$$s_x = h s_u = 5 \sqrt{\frac{193}{75} - \left(\frac{-9}{75}\right)^2} = 5 \times \sqrt{2.57 - .0144}$$

$$= 5 \times 1.598 = 7.99 \text{ (app).}$$

**Root Mean Square Deviation :** When the deviations of the observations are taken from any arbitrary value A other than  $\bar{x}$ , standard deviation reduces to root mean square deviation which is defined by

$$s' = \sqrt{\frac{1}{n} \sum_{i=1}^k f_i (x_i - A)^2} \quad \dots \dots \dots (4.9)$$

where A is an arbitrary value and  $n = \sum_{i=1}^k f_i$ .



## Measures of Dispersion

**Theorem 4.2** Standard deviation is the least possible root mean square deviation i. e. root mean square deviation is the least when the deviations are taken from the arithmetic mean.

**Proof :** Let  $x_1, x_2, \dots, x_k$  are the values of  $k$  observations with corresponding frequencies  $f_1, f_2, \dots, f_k$ . Also let  $\bar{x}$  be the mean of the observations and  $A$  be any arbitrary value.

$$\text{We know, } s'^2 = \frac{1}{n} \sum f_i (x_i - A)^2 = \frac{1}{n} \sum f_i \{(x_i - \bar{x}) + (\bar{x} - A)\}^2$$

$$= \frac{1}{n} \left[ \sum f_i (x_i - \bar{x})^2 + 2(\bar{x} - A) \sum f_i (x_i - \bar{x}) + n(\bar{x} - A)^2 \right]$$

$$= \frac{1}{n} \sum f_i (x_i - \bar{x})^2 + 0 + \text{a positive value}$$

$$= s^2 + \text{a positive value}$$

Therefore,  $s'^2 > s^2$  i. e.  $s' > s$  ... .. (4.10)

Hence the theorem is proved.

**Theorem 4.3** For two observations, standard deviation is the half of the range.

**Proof :** Let  $x_1$  and  $x_2$  be two observations. Then  $\bar{x} = \frac{(x_1 + x_2)}{2}$ ,

where  $\bar{x}$  is the arithmetic mean. Let  $s$  denote the standard deviation.

$$\text{From (4.2) we have, } s^2 = \frac{1}{2} \left[ \left\{ x_1 - \frac{(x_1 + x_2)}{2} \right\}^2 + \left\{ x_2 - \frac{(x_1 + x_2)}{2} \right\}^2 \right]$$

$$= \frac{1}{2} \left\{ \left( \frac{x_1 - x_2}{2} \right)^2 + \left( \frac{x_2 - x_1}{2} \right)^2 \right\} = \left( \frac{x_1 - x_2}{2} \right)^2$$

Therefore,  $s = \left| \frac{x_1 - x_2}{2} \right|$   $\therefore$  Hence proved.

**Theorem 4.4** The standard deviation of first  $n$  natural numbers is  $\sqrt{\frac{(n^2 - 1)}{12}}$

**Proof:** The  $n$  natural numbers are  $1, 2, 3, \dots, n$ .


$$\text{Their mean, } \bar{x} = \frac{1 + 2 + \dots + n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\text{The raw sum of squares, } \sum x_i^2 = 1^2 + 2^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\text{We know, } s^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{n(n+1)(2n+1)}{6n} - \left\{ \frac{(n+1)}{2} \right\}^2$$

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

$$= \frac{(n+1)(n-1)}{2 \times 6} = \frac{n^2-1}{12}$$

Therefore,  $s = \sqrt{\frac{(n^2-1)}{12}}$ . Hence proved. 

**Theorem 4.5** If  $\bar{x}$  and  $s$  be arithmetic mean and standard deviation respectively of  $n$  non-negative observations, then  $\bar{x} \sqrt{(n-1)} \geq s$ .

**Proof:** Let us consider  $x_1, x_2, \dots, x_n$  be  $n$  non-negative observations.

$$\text{We know } \sum x_i = n \bar{x} \text{ and } ns^2 = \sum x_i^2 - \frac{\left( \sum x_i \right)^2}{n}$$

$$\text{Now } \left( \sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n x_i^2 + \sum_{i \neq j} x_i x_j$$

Since  $x_i$ 's are non-negative,  $\sum_{i \neq j} x_i x_j \geq 0$

Hence,  $\left( \sum_{i=1}^n x_i \right)^2 \geq \sum_{i=1}^n x_i^2$ . Subtracting  $\frac{\left( \sum x_i \right)^2}{n}$  from both sides.

we have,  $\left(\frac{n}{\sum x_i}\right)^2 - \frac{\left(\sum x_i\right)^2}{n} \geq \frac{n}{\sum x_i^2} - \frac{\left(\sum x_i\right)^2}{n}$

or,  $\frac{(n-1)}{n} \left(\frac{n}{\sum x_i}\right)^2 \geq ns^2$

or,  $\frac{(n-1)}{n} (n \bar{x})^2 \geq ns^2$

or,  $(n-1) \bar{x}^2 \geq s^2$

or,  $\sqrt{(n-1) \bar{x}^2} \geq s$ , Hence proved

**Theorem 4.6** (Standard deviation of combined series) If  $n_1$  and  $n_2$  be the number of observations,  $\bar{x}_1$  and  $\bar{x}_2$  be the means and  $s_1$  and  $s_2$  be the standard deviations of two series, then the standard deviation  $s$  of the combined series is given by,

$$s = \left[ \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^2 \right]^{\frac{1}{2}}$$

**Proof :** Let  $x_{1i}$  ( $i = 1, 2, \dots, n_1$ ) and  $x_{2j}$  ( $j = 1, 2, \dots, n_2$ ) are two series with means  $\bar{x}_1$  and  $\bar{x}_2$  and variances  $s_1^2$  and  $s_2^2$  respectively. The mean of the combined series is

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Let  $d_1 = \bar{x}_1 - \bar{x}$  and  $d_2 = \bar{x}_2 - \bar{x}$ .

The variance of the combined series is given by

$$s^2 = \frac{1}{n_1 + n_2} \left[ \sum_{i=1}^{n_1} (x_{1i} - \bar{x})^2 + \sum_{j=1}^{n_2} (x_{2j} - \bar{x})^2 \right] \dots \dots (4.11)$$

Now,  $\sum_{i=1}^{n_1} (x_{1i} - \bar{x})^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1 + \bar{x}_1 - \bar{x})^2$



$$= \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + n_1 (\bar{x}_1 - \bar{x})^2 = n_1 s_1^2 + n_1 d_1^2 \quad \dots \dots (4.12)$$

Similarly we get,  $\sum_{j=1}^{n_2} (x_{2j} - \bar{x})^2 = n_2 s_2^2 + n_2 d_2^2 \quad \dots (4.13)$

Substituting (4.12) and (4.13) in (4.11), we have

$$s^2 = \frac{1}{n_1 + n_2} [n_1 s_1^2 + n_2 s_2^2 + n_1 d_1^2 + n_2 d_2^2] \quad \dots (4.14)$$

We know,  $d_1 = \bar{x}_1 - \bar{x} = \bar{x}_1 - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{n_2 (\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}$

and  $d_2 = \bar{x}_2 - \bar{x} = \frac{n_1 (\bar{x}_2 - \bar{x}_1)}{n_1 + n_2}$

Putting the values of  $d_1$  and  $d_2$  in (4.14) and after simplification, we get,

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^2$$

$$\therefore s = \left[ \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^2 \right]^{1/2} \quad \dots (4.15)$$

**Remark :** We can generalise the result in (4.14) to get the combined standard deviation of  $k$  series as

$$s = \left[ \sum_{i=1}^k \frac{n_i}{n} (s_i^2 + d_i^2) \right]^{1/2} \quad \dots \dots \dots (4.16)$$

where  $d_i = \bar{x}_i - \bar{x}$ ;  $i = 1, 2, \dots, k$ ;  $n = n_1 + n_2 + \dots + n_k$

and  $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n}$

**Example 4.3** A group of 40 students was selected and measured their heights which give mean 4.5 ft. and standard deviation 2.1 ft. Another group of 50 students was also selected and measured their heights whose mean is 4.3 ft. and standard deviation is 1.9 ft. Find the mean and standard deviation of the combined group.

**Solution :** We have,

Group	No. of sample	Mean	Variance
1st	$n_1 = 40$	$\bar{x}_1 = 4.5$	$s_1^2 = 4.41$
2nd	$n_2 = 50$	$\bar{x}_2 = 4.3$	$s_2^2 = 3.61$

## Measures of Dispersion

From (4.15) we have,  $s = \left[ \frac{176.4 + 180.5}{90} + \frac{2000 \times 0.04}{8100} \right]^{\frac{1}{2}} = 2.016$  ft.

**Example 4.4** A student while calculating mean and standard deviation of 25 observations obtained mean = 56 and standard deviation = 2. At the time of checking it was found that he has copied 64 instead of 46. What would be the actual value of mean and standard deviation?

**Solution:** We have,  $n = 25$ ,  $\bar{x} = 56$  and  $s = 2$

We know,  $\sum x_i = n \bar{x} = 25 \times 56 = 1400$ .

Since the observation 64 was wrongly included instead of 46, the  $\sum x_i$  would be  $1400 - 64 + 46 = 1382$ .

Therefore, the actual mean,  $\bar{x}' = \frac{1382}{25} = 55.28$ .

Also we know,  $s^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2$

or,  $\sum x_i^2 = n(s^2 + \bar{x}^2) = 25(2^2 + 56^2) = 78500$

which is the sum of squares of the observations considering 64.

Therefore, the actual raw sum of squares would be

$\sum x_i^2 = 78500 - 64^2 + 46^2 = 76520$ .

Therefore, the actual standard deviation

$s' = \left( \frac{76520}{25} - (55.28)^2 \right)^{\frac{1}{2}} = (3060.80 - 3055.88)^{\frac{1}{2}} = 2.2$  (app)

Thus the actual mean = 55.28 and the standard deviation = 2.2 (app)

### Merits :

- 1) It is rigidly defined.
- 2) It takes all the observations into account.
- 3) It is amenable to algebraic treatments.
- 4) It is less affected by sampling fluctuation.
- 5) The standard deviation of the combined series can be obtained if the means, standard deviations and number of observations in each series are given.

### Demerits :

- 1) It is not readily comprehensible, the calculation requires a good deal of time and knowledge of arithmetic.
- 2) It is affected by the extreme values.

- 3) It cannot be calculated if the extreme classes of the frequency distribution are open.

**Uses of Standard Deviation :** It is the most useful measure of dispersion and has got immense use in advanced statistical work such as sampling, correlation analysis, normal curve of errors, comparing variability and uniformity of two sets of data etc.

**4. Mean Deviation :** It would be useless to take the sum of the deviations of the values from the arithmetic mean as a measure of dispersion since their algebraic sum is zero.

However, if we take the mean of the absolute values of the deviations we get a useful measure of dispersion called mean deviation from mean, or mean absolute deviation or simply mean deviation.

Let  $x_1, x_2, \dots, x_n$  be  $n$  observations of a variable with mean  $\bar{x}$ , then mean deviation is defined by

$$M. D (\bar{x}) = \frac{1}{n} \sum |x_i - \bar{x}| \quad \dots \dots (4.17)$$

$$\text{In case of grouped data, } M. D (\bar{x}) = \frac{1}{n} \sum f_i |x_i - \bar{x}| \quad \dots \dots (4.18)$$

If the deviations are taken from median or mode,  $\bar{x}$  has to be replaced from  $k$

(4.17) and (4.18) and then for grouped data with  $\sum f_i = n$

$$M. D (Me) = \frac{1}{n} \sum f_i |x_i - Me| \quad \dots \dots (4.19)$$

$$\text{and } M. D (Mo) = \frac{1}{n} \sum f_i |x_i - Mo| \quad \dots \dots (4.20)$$

Quiz  
✓

**Example 4.5** Calculate mean deviation from the mean and mean deviation from the median from the frequency distribution given in Example 4.1.

**Solution :** We know that arithmetic mean or simply mean of the distribution is 21.9, and the median is 22.5. We prepare Table - 4.2 for calculation of mean deviation from both mean and median.



## Measures of Dispersion

Table -4.2

Class Int. of variable	Mid values $x_i$	Frequency $f_i$	$ x_i - \bar{x} $	$f_i  x_i - \bar{x} $	$ x_i - Me $	$f_i  x_i - Me $
0—5	2.5	2	19.4	38.8	20.0	40
5—10	7.5	5	14.4	72.0	15.0	75
10—15	12.5	7	9.4	65.8	10.0	70
15—20	17.5	13	4.4	57.2	5.0	65
20—25	22.5	21	.6	12.6	0	0
25—30	27.5	16	5.6	89.6	5	80
30—35	32.5	8	10.6	84.8	10	80
35—40	37.5	3	15.6	46.8	15	45
Total		75		467.6		455

Therefore,  $M. D (\bar{x}) = \frac{1}{n} \sum_{i=1}^k f_i |x_i - \bar{x}| = \frac{467.6}{75} = 6.23$  (app)

and  $M. D (Me) = \frac{1}{n} \sum_{i=1}^k f_i |x_i - Me| = \frac{455}{75} = 6.07$  (app)

**Theorem 4.7** Mean deviation is the least when the deviations are measured from median.

**Proof :** Let  $n = 2p$  be the number of observations which are arranged in order of magnitude as  $x_1, x_2, \dots, x_p, x_{p+1}, \dots, x_n$ . Hence the median  $Me$  lies between  $x_p$  and  $x_{p+1}$ . Let  $A$  be an arbitrary value less than  $Me$ , and lies between  $x_k$  and  $x_{k+1}$  ( $k < p$ ).

Considering absolute deviations  $D_1$  and  $D_2$  from  $Me$  and  $A$  respectively we have,

$$\begin{aligned}
 D_1 &= (Me - x_1) + (Me - x_2) + \dots + (Me - x_k) \\
 &\quad + (Me - x_{k+1}) + (Me - x_{k+2}) + \dots + (Me - x_p) \\
 &\quad + (x_{p+1} - Me) + (x_{p+2} - Me) + \dots + (x_n - Me) \quad \dots \dots \dots (4.21)
 \end{aligned}$$

$$\begin{aligned}
 D_2 &= (A - x_1) + (A - x_2) + \dots + (A - x_k) \\
 &\quad + (x_{k+1} - A) + (x_{k+2} - A) + \dots + (x_p - A) \\
 &\quad + (x_{p+1} - A) + (x_{p+2} - A) + \dots + (x_n - A) \quad \dots \dots \dots (4.22)
 \end{aligned}$$

(4.22) - (4.21) gives  $D_2 - D_1 = (A - Me) k - (p - k) (A + Me) + 2(x_{k+1} + x_{k+2} + \dots + x_p) + p (Me - A)$ .

$$= 2[(x_{k+1} - A) + (x_{k+2} - A) + \dots + (x_p - A)]$$

which is a positive quantity. Hence the theorem is proved.

The same proof can be done if we consider  $Me < A$  and also if the number of observations are odd.

**Remark :** In some practical situation the theorem may contradict as the value of median cannot be obtained exactly in grouped frequency distribution.

**Theorem 4.8** For a series of  $n$  observations, standard deviation is not less than the mean deviation from mean.

**Proof :** Let  $x_1, x_2, \dots, x_k$  be a series of observations with corresponding frequencies  $f_1, f_2, \dots, f_k$  such that  $\sum f_i = n$ . We have to prove that

$s \geq M.D(\bar{x})$  where  $s$  = standard deviation. or,  $s^2 \geq [M.D(\bar{x})]^2$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^k f_i (x_i - \bar{x})^2 \geq \left[ \frac{1}{n} \sum_{i=1}^k f_i |x_i - \bar{x}| \right]^2 \quad \dots \dots (4.23)$$

Let us put  $x_i - \bar{x} = z_i$  in (4.23) we have

$$\frac{1}{n} \sum_{i=1}^k f_i z_i^2 \geq \left( \frac{1}{n} \sum_{i=1}^k f_i z_i \right)^2$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^k f_i z_i^2 - \bar{z}^2 \geq 0$$

or,  $\frac{1}{n} \sum_{i=1}^k f_i (z_i - \bar{z})^2 \geq 0$ . Hence proved.

**Example 4.6** Find the mean deviation from mean and standard deviation of the observations like  $a, a + d, a + 2d, \dots, a + 2nd$ . And prove that the latter is greater than the former.

**Solution :** The number of observations are  $2n + 1$ .

The arithmetic mean,  $\bar{x} = \frac{a + (a + d) + \dots + (a + 2nd)}{2n + 1}$

### Measures of Dispersion

$$= \frac{\{(2n+1)a + d(1+2+\dots+\dots+2n)\}}{2n+1}$$

$$= \frac{\{(2n+1)a + dn(2n+1)\}}{2n+1} = a + nd$$

The absolute deviations of the observations from the mean are,  $nd, (n-1)d, (n-2)d, \dots, (n-2)d, (n-1)d, nd$ . and their sum is equal to

$$d[n + (n-1) + (n-2) + \dots + 1 + 1 + \dots + (n-2) + (n-1) + n]$$

$$= 2d[n + (n-1) + (n-2) + \dots + 1] \text{ since each term occurs twice,}$$

$$= \frac{2nd(n+1)}{2} = dn(n+1)$$

$$\therefore \text{Mean deviation from mean} = \frac{dn(n+1)}{2n+1}$$

$$\text{Now, variance} = \frac{1}{2n+1} [n^2 d^2 + (n-1)^2 d^2 + \dots + (n-1)^2 d^2 + n^2 d^2]$$

$$= \frac{2d^2}{2n+1} [n^2 + (n-1)^2 + \dots + 1^2]$$

$$= \frac{2d^2}{2n+1} \times \frac{n(n+1)(2n+1)}{6} = \frac{d^2 n(n+1)}{3}$$

$$\therefore \text{Standard deviation} = d \sqrt{\frac{n(n+1)}{3}}$$

We are to show that the standard deviation is greater than mean deviation.

$$\text{That is, squaring, } \frac{d^2 n(n+1)}{3} > \frac{d^2 n^2 (n+1)^2}{(2n+1)^2}$$

$$\text{or, } (2n+1)^2 > 3n(n+1)$$

$$\text{or, } n^2 + n + 1 > 0, \text{ which is true, Hence the result.}$$

**Merits :**

- 1) It is easy to understand.
- 2) It is relatively easy to calculate.
- 3) It takes all the observations into account.
- 4) It is less affected by the extreme values.



**Demerits :**

- 1) It is not amenable to further algebraic treatments.
- 2) It cannot be calculated if the extreme classes of the frequency distribution are open.
- 3) It is less stable than standard deviation.

**Remark :** Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations as  $|x_i - \bar{x}|$  creates artificiality and renders it useless for further mathematical treatments.

**4.4 Empirical Relation Among Quartile Deviation, Standard Deviation and Mean Deviation**

For moderately asymmetrical distribution the following empirical relations hold approximately :

- 1) Quartile deviation =  $\frac{2}{3}$  Standard deviation.
- 2) Mean deviation from mean =  $\frac{4}{5}$  Standard deviation.

**Relative Measures of Dispersion**

Whenever we want to compare the variability of two series which differ wide in their measures of central location or which are measured in different units, the absolute measures of dispersion donot serve our purpose properly. To tackle this situation we usually calculate the relative measures of dispersion which are pure numbers, independent of units of measurement.

**1. Co-efficient of Quartile Deviation :** The co-efficient of quartile deviation is defined by

$$C. Q. D = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 \quad \dots \dots (4.24)$$

where  $Q_3$  is the third quartile and  $Q_1$  is the first quartile.

**Example 4.7** Find co-efficient of quartile deviation from the frequency distribution given in Example 4.1.

**Solution :** We know,  $Q_3 = 27.58$  and  $Q_1 = 16.83$ .

Therefore,  $C. Q. D = \frac{10.75}{44.41} \times 100 = 24.21$  (app)

## Measures of Dispersion

**2. Co-efficient of Variation :** The co-efficient of variation is defined by

$$C. V = \frac{s}{\bar{x}} \times 100 \quad \dots \dots \dots (4.25)$$

where  $s$  is the standard deviation and  $\bar{x}$  is the mean. Co-efficient of variation is, however, unreliable if  $\bar{x}$  is near to zero. It can be easily shown that co-efficient of variation is independent on change of scale but not of origin.

**Example 4.8** Calculate co-efficient of variation from the frequency distribution given in Example 4.1.

**Solution :** We know that  $s = 7.99$  (app) and  $\bar{x} = 21.9$  (app)

Therefore,  $C. V = \frac{7.99}{21.9} \times 100 = 36.48\%$  (app).

**3. Co-efficient of Mean Deviation :** The co-efficient of mean deviation (C. M. D.) is the ratio of the mean deviation measured from certain measure of central location to the corresponding measure of central location and is expressed as percentage.

$$\left. \begin{aligned} \text{That is, a) C. M. D. } (\bar{x}) &= \frac{M. D. (\bar{x})}{\bar{x}} \times 100 \\ \text{b) C. M. D. (Me)} &= \frac{M. D. (Me)}{Me} \times 100 \\ \text{c) C. M. D. (Mo)} &= \frac{M. D. (Mo)}{Mo} \times 100 \end{aligned} \right\} \dots (4.26).$$

**Example 4.9** Calculate co-efficient of mean deviation from median from the frequency distribution given in Example 4.1.

**Solution :** We know  $M. D. (Me) = 6.07$  and  $Me = 22.5$

Therefore,  $C. M. D. (Me) = \frac{6.07}{22.5} \times 100 = 26.98\%$  (app)

### 4.5 Moments

If  $x_1, x_2, \dots, x_n$  be  $n$  observations of a variate then the  $r$ th raw moment is defined by

$$\mu'_r = \frac{1}{n} \sum (x_i - A)^r; \text{ where } A \text{ is any arbitrary value} \quad (4.27)$$

The rth corrected moment is defined by

$$\mu_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r \text{ where } \bar{x} \text{ is arithmetic mean} \quad \dots (4.28)$$

If  $x_1, x_2, \dots, x_k$  occur with frequencies  $f_1, f_2, \dots, f_k$  respectively then the rth raw moment is

$$\mu'_r = \frac{1}{n} \sum_{i=1}^k f_i (x_i - A)^r; \text{ where } n = \sum_{i=1}^k f_i \text{ and } A \text{ is as earlier} \quad \dots (4.29)$$

The rth corrected moment is defined by

$$\mu_r = \frac{1}{n} \sum_{i=1}^k f_i (x_i - \bar{x})^r, \text{ where } n = \sum_{i=1}^k f_i, \text{ and } \bar{x} = \text{arithmetic mean.} \quad (4.30)$$

**Relation Between Raw Moments and Corrected Moments :**

$$\text{We have, } \mu_r = \frac{1}{n} \sum_{i=1}^k f_i (x_i - \bar{x})^r = \frac{1}{n} \sum_{i=1}^k f_i \{ (x_i - A) - (\bar{x} - A) \}^r$$

$$= \frac{1}{n} \left[ \sum_{i=1}^k f_i (x_i - A)^r - \binom{r}{1} d \sum_{i=1}^k f_i (x_i - A)^{r-1} + \right.$$

$$\left. \binom{r}{2} d^2 \sum_{i=1}^k f_i (x_i - A)^{r-2} - \dots + (-1)^r d^r \right] \quad \dots (4.31)$$

where  $d = \bar{x} - A$ . Now according to the definitions given earlier we have,

$$\mu_r = \mu'_r - \binom{r}{1} d \mu'_{r-1} + \binom{r}{2} d^2 \mu'_{r-2} - \binom{r}{3} d^3 \mu'_{r-3} + \dots + (-1)^r d^r \mu'_0 \quad \dots (4.32)$$

$$\text{we know, } \mu'_1 = \frac{1}{n} \sum_{i=1}^k f_i (x_i - A) = \bar{x} - A = d.$$

Now putting  $r = 2, 3, 4, \dots$  etc; and  $d = \mu'_1$  in (4.32) we have,



## Measures of Dispersion

$$\left. \begin{aligned}
 \mu_2 &= \mu'_2 - 2\mu'_1\mu'_1 + \mu'_1{}^2 = \mu'_2 - \mu'_1{}^2 \\
 \mu_3 &= \mu'_3 - 3\mu'_1\mu'_2 + 3\mu'_1{}^2\mu'_1 - \mu'_1{}^3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'_1{}^3 \\
 \mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_1{}^2\mu'_2 - 4\mu'_1{}^3\mu'_1 + \mu'_1{}^4 \\
 &= \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_1{}^2\mu'_2 - 3\mu'_1{}^4
 \end{aligned} \right\} \dots\dots\dots(4.33)$$

Using the method given in Theorem 4.1, it can be easily shown that the moments are independent on change of origin but not of scale.

**Sheppard's Correction for Grouping :** It is generally assumed that the frequencies in a group are concentrated at the mid point of the class-interval. This is surly an approximation. W. F. Sheppard observed that if :

- a) the frequency distribution is continuous and
- b) the frequency tappers off to zero in both directions of the frequency distribution,

then the correction for different moments due to grouping at the mid point of the class interval are done by the following formula. This is known as Sheppard's correction.

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{c^2}{12}$$

$$\mu_3 \text{ (Corrected)} = \mu_3$$

$$\mu_4 \text{ (Corrected)} = \mu_4 - \frac{1}{2} c^2 \mu_2 + \frac{7}{240} c^4, \quad \text{where } c \text{ is the length of class interval.}$$

### Pearson's $\beta$ and $\gamma$ Co-efficients :

Karl Pearson defined the following co-efficients, based on the first four corrected moments :

$$\beta_1 = \frac{\mu_3}{\mu_2^2}; \quad \gamma_1 = +\sqrt{\beta_1}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^3}; \quad \gamma_2 = \beta_2 - 3.$$

The above co-efficients are pure numbers and independent of units of measurement. Practical utility of these co-efficients is discussed in section 4.6.

**Theorem 4.8** Pearson's  $\beta$  co-efficients satisfy the following inequalities.

- i)  $\beta_2 \geq \beta_1 + 1,$  ii)  $\beta_2 \geq 1.$

**Proof :** Let  $x_1, x_2, \dots, x_n$  be a set of  $n$  observations with mean  $\bar{x}$  i. e.  $\sum x_i = n\bar{x}$

We know,  $\sum_{i=1}^n (ax_i^2 + bx_i + c)^2 \geq 0$ ; where a, b and c are arbitrary constants.

Expanding the left hand side we have,

$$\sum_{i=1}^n (a^2x_i^4 + b^2x_i^2 + c^2 + 2abx_i^3 + 2acx_i^2 + 2bcx_i) \geq 0$$

$$\text{or, } a^2 \sum_{i=1}^n x_i^4 + b^2 \sum_{i=1}^n x_i^2 + nc^2 + 2ab \sum_{i=1}^n x_i^3 + 2ac \sum_{i=1}^n x_i^2 \geq 0$$

Dividing by n we have,

$$a^2\mu_4 + b^2\mu_2 + c^2 + 2ab\mu_3 + 2ac\mu_2 \geq 0, \text{ since } \bar{x} = 0.$$

Putting  $a = 1, b = \frac{-\mu_3}{\mu_2}$  and  $c = -\mu_2$  and dividing by  $\mu_2^2$  we have,

$$\frac{\mu_4}{\mu_2^2} + \frac{\mu_3^2}{\mu_2^3} + 1 - 2\frac{\mu_3^2}{\mu_2^3} - 2 \geq 0.$$

$$\text{or, } \beta_2 + \beta_1 + 1 - 2\beta_1 - 2 \geq 0 \text{ or, } \beta_2 - \beta_1 - 1 \geq 0 \text{ or, } \beta_2 \geq \beta_1 + 1.$$

Since  $\beta_1 \geq 0$ ; we get  $\beta_2 \geq 1$ . Hence proved.

## 1.6 Skewness and Kurtosis

**Skewness** : Skewness means "lack of symmetry" i. e. departure from symmetry of a distribution. A distribution is said to be skewed if :

- 1) Mean, Median and Mode give different values.
- 2)  $Q_1$  and  $Q_3$  are not equidistant from median
- 3) The curve drawn with the help of the given data is not symmetrical but turned nose to one side than the other. If the curve has a longer tail to the right side, then the distribution is said to be positively skewed and in the reverse case it is negatively skew.

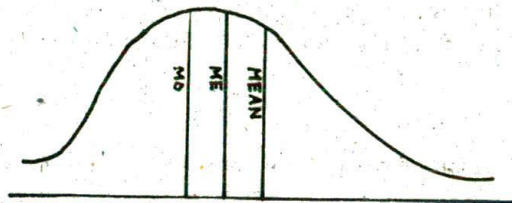


Fig. 4.2 (a) Positive skew curve

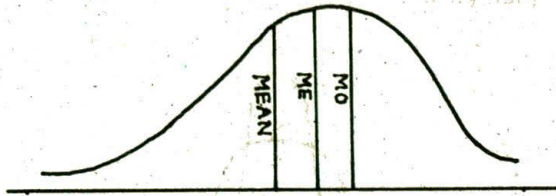


Fig 4.2 (b) Negative skew curve.

**Measures of Skewness :** Different absolute measures of skewness are

- 1) S. K. = 3 (Mean — Median). 2) S. K. = Mean — Mode
- 3) S. K. =  $(Q_3 - Q_2) - (Q_2 - Q_1) = Q_3 + Q_1 - 2Q_2$

where  $Q_1$ ,  $Q_2$  and  $Q_3$  are the first, second and third quartiles respectively.

Different relative measures of skewness are called co-efficient of skewness which are pure numbers and independent on units of measurement. The following are the co-efficients of skewness.

- 1) Prof. Karl Pearson's co-efficient of skewness is given by

$$C. S. K. = \frac{\text{Mean} - \text{Mode}}{\text{st. deviation}} \text{ and } C. S. K. = \frac{3(\text{Mean} - \text{Median})}{\text{st. deviation}}$$

- 2) Prof. Bowlays' Co-efficient of skewness is given by

$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

- 3) Co-efficient of Skewness based on moments is given by

$$\frac{\sqrt{\beta_1 (\beta_2 + 3)}}{2(5\beta_2 - 6\beta_1 - 9)} \text{ where } \beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

We know that,  $\mu_2$ ,  $\mu_3$  and  $\mu_4$  are the second, third and fourth corrected moments respectively.

**Kurtosis :** The degree of peakness or flatness of a distribution relative to a normal distribution (discussed in chapter 8) is called kurtosis.

The measure of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \gamma_2 = \beta_2 - 3 \text{ is called the excess of kurtosis or simply excess.}$$



For normal distribution,  $\beta_2 = 3$  i. e.  $\gamma_2 = 0$  the curve is called **mesokurtic**. When  $\beta_2 > 3$  i. e.  $\gamma_2 > 0$ , the curve is called **leptokurtic** and when  $\beta_2 < 3$  i. e.  $\gamma_2 < 0$ , the curve is called **platykurtic**.

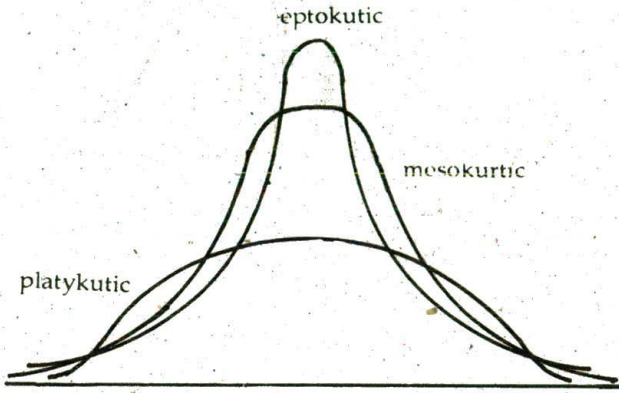


Fig. 4.3 Different types of kurtosis

## 5. THEORY OF PROBABILITY

### 5.1 Introduction

In our everyday life, we used to comment on so many occasions like, there is 60% chance of raining to-morrow, the chance of winning of a particular team in a football match is very high, the chance of success of a particular student is very little etc. In all these statements we have an idea of probability in our mind.

Usually we draw conclusion of a certain characteristics about the population on the basis of sample. The conclusion is bound to be uncertain to a greater or lesser extent. The notion of uncertainty plays a vital role in statistics and the measurement of the degree of uncertainty in a decision or conclusion is naturally of prime importance. The theory of probability mainly deals in this measurement. The mathematical theory of probability was laid in the mid-seventeenth century based on the problem of game theory.

### 5.2 Definitions of Various Terms

(a) **Experiment** : An experiment is an act that can be repeated under given conditions.

(b) **Trial and Event** : If an experiment be repeated under essentially the same condition giving several possible outcomes then the experiment is called a trial and the possible outcomes are known as events or cases.

For example, tossing of a coin is a trial and getting Head (H) or Tail (T) is an event. The trial which contains no possible event is called impossible event. The probability of an impossible event is 0. The trial which contains all the possible cases or events is called a sure event. The probability of a sure event is 1.

(c) **Exhaustive Cases** : The total number of possible outcomes of any 'trial' are exhaustive cases. For example, in tossing a coin there are two exhaustive cases namely occurrence of head and of tail.

(d) **Equally Likely Cases** : Cases are said to be equally likely when none of them is expected to occur more frequently than the other. For example, from an unbiased coin, the case of appearing head or tail is equally likely.

(e) **Mutually Exclusive Cases** : Cases are said to be mutually exclusive if the occurrence of one of them excludes the occurrence of all the others. For example, in tossing an unbiased coin, the cases of appearing head and tail are mutually exclusive.

(f) **Favourable Cases** : The number of outcomes which entail the happening of the event in a trial is called the favourable cases. For example, the number of favourable cases for getting even numbers in tossing a die is 3.

### 5.3 Definitions of Probability

There are mainly two definitions of probability, namely

- (i) Mathematical or classical or a-priori definition of probability.
- (ii) Statistical or empirical or a-posteriori definition of probability.

**Mathematical or Classical or a-priori definition of Probability** : If a trial results in  $n$  exhaustive, mutually exclusive and equally likely cases and  $m$  of them are favourable to an event  $A$ , then the probability  $p$  of the happening of  $A$  is given by

$$p = \frac{\text{Favourable number of cases}}{\text{Total number of cases}} = \frac{m}{n}$$

This gives the numerical measure of probability. Obviously  $p$  be a positive number not greater than unity, so that  $0 \leq p \leq 1$ .

Since the number of cases in which the event  $A$  will not happen is  $(n - m)$ , the probability  $q$  that the event will not happen is given by  $q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p$ . So that  $p+q=1$ .

**Remarks : 1.** The probability of the happening of an event is known as probability of success and the probability of the non-happening of the event is the probability of failure.

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

#### Limitations of classical definition of probability :

We cannot define this kind of probability if

- i) the outcomes are not equally likely and
- ii) the number of outcomes are infinite.



**Example 5.1** A bag contains 15 identical balls of which 5 are white and the rest are black. Two balls are drawn at random from the bag. What is the probability that both balls are white?

**Solution :** Let  $A$  be the event that both the balls are white. The total number of cases of getting 2 white balls from 15 balls is  ${}^{15}C_2 = \frac{14 \times 15}{2} = 105$ . The favourable number of cases of getting 2 white balls from 5 white balls is  ${}^5C_2 = \frac{4 \times 5}{2} = 10$ .

Therefore, the required probability is  $P(A) = \frac{10}{105} = \frac{2}{21}$

**Statistical or Empirical or a-posteriori definition of probability :** If a trial is repeated a number of times essentially under the same condition then the limit of the ratio of the number of times that an event happens to the total number of trials as the number of trials increase indefinitely is called the probability of the happening of that event. It is assumed that the limit is finite and unique.

Symbolically, if  $n$  be the number of cases of a trial  $A$  and  $m$  be the number of cases that the trial  $A$  can happen then the probability  $p$  of the happening of the trial  $A$  is given by

$$p = \lim_{n \rightarrow \infty} \frac{m}{n}$$

For example, if an unbiased coin is tossed in the following number of times and the number of times of getting head upward are recorded against each experiment, then we can see that as the number of tosses increase the probability of getting head upward approaches to the true probability.

**Table -5.1**

Number of tosses	Number of heads obtained	Probability of getting head
10	4	.40
20	9	.45
50	26	.52
100	51	.51
1000	500	.50

#### 5.4 Elementary Set Theory (Notations, Operations and Algebra)

**Set :** A set is any well defined list, collection or class of objects having common quality or feature. The objects in a set is called the elements or members of the set.

Following are the examples of set :

- i) The numbers 2, 4, 6 and 8.
- ii) The days of the week
- iii) The solutions of the equation,  $x^2-3x+2=0$ .

Sets are usually denoted by capital letters A, B, C, X, Y etc., if x is an element of the set A, we write symbolically  $x \in A$  (x belongs to A). If x is not a member of A, then we write  $x \notin A$  (x does not belong to A). A set is written by the elements which are separated by commas enclosed in brackets { }.

**Finite and Infinite Set :** A set is finite if it consists of specific number of different elements, otherwise the set is infinite. Let M be the set of the days in a week then it is finite. Again let  $N=\{2, 4, 6, \dots\}$  then it is an infinite set.

**Equality of sets :** Set A is equal to set B if they both have the same elements, i. e. if every element which belongs to A also belongs to B and if every element which belongs to B also belongs to A. We denote the equality of sets A and B by  $A=B$ .

**Null Set :** A set which contains no element is called null set. We denote it by the symbol  $\Phi$ .

**Sub Set :** If every element in a set A is also a member of a set B then A is called a sub set of B. We denote the relationship by writing  $A \subset B$ , which can also be read "A is contained in B".

**Disjoint Set :** If set A and B have no elements in common i. e. if no element of A is in B and no element of B is in A, then we say that A and B are disjoint.

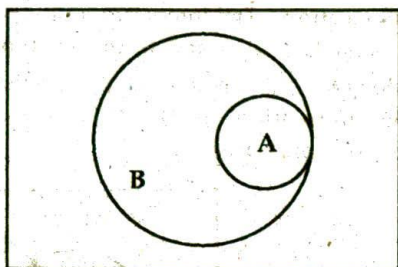
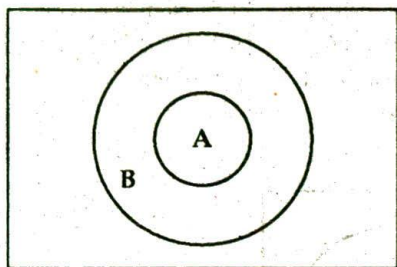
For example, if  $A=\{1, 3, 4\}$  and  $B=\{2, 5, 7\}$ , then A and B are disjoint.

**Comparability :** Two sets are said to be comparable if  $A \subset B$  or  $B \subset A$ , i. e. if one of the sets is a subset of the other. Moreover, two sets A and B are said to be not comparable if  $A \not\subset B$  or  $B \not\subset A$ .

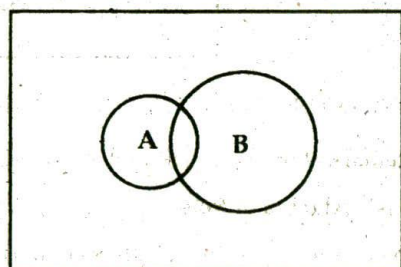
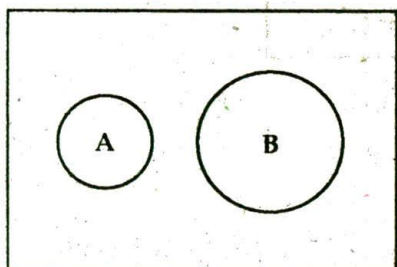
**Venn Diagram :** A simple and instructive way of representing the relationship between sets is given in Venn diagrams. We can represent a set by a simple plane area usually bounded by a circle.

## Theory of Probability

Let us suppose that the sets A and B are such that  $A \subset B$  and  $A \neq B$ , then A and B can be represented by either of the following diagram.



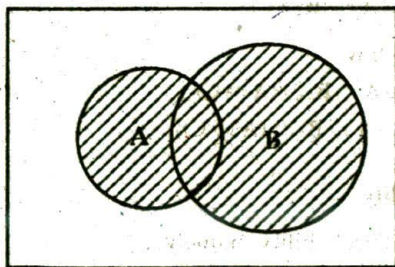
Again let us suppose that A and B are not comparable. Then A and B are disjoint when they follow the left diagram and are not disjoint when they follow the right diagram.



### Basic Set Operations :

**Union :** The union of sets A and B is the set of all elements which belongs to A or to B or both. We denote the union of A and B by  $A \cup B$ , which is usually read "A union B". It follows directly from the definition of the union of two sets that  $A \cup B$  and  $B \cup A$  are the same set i. e.  $A \cup B = B \cup A$ .

In the Venn diagram we have shaded  $A \cup B$

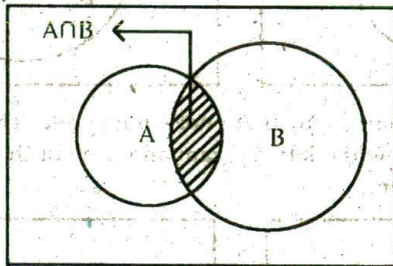




## An Introduction to The Theory of Statistics

For example ; Let  $A = \{2, 3, 5\}$  and  $B = \{2, 3, 4, 6\}$ , then  $A \cup B = \{2, 3, 4, 5, 6\}$

**Intersection :** The intersection of sets A and B is the set of elements which are common to A and B, that is, those elements which belong to A also belong to B. We denote the intersection of A and B by  $A \cap B$ , which is read "A intersection of B." In the Venn-diagram we have shaded the intersectional area.



For example : Let  $A = \{2, 3, 5\}$  and  $B = \{2, 3, 4, 6\}$ , then  $A \cap B = \{2, 3\}$ .

**Remark :** For two disjoint sets A and B;  $A \cap B = \Phi$ .

### Basic Algebra of Sets

In the theory of sets, all the sets under investigation will likely be sub-sets of a fixed set. We shall call this set a universal set.

Let A, B and C are the sub-sets of a universal set, then the following laws hold :

i) **Commutative law :**

$$A \cup B = B \cup A ; A \cap B = B \cap A.$$

ii) **Associative law :**

$$(A \cup B) \cup C = A \cup (B \cup C).$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

iii) **Distributive law :**

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

### 5.5 Laws of Probability

There are two laws of probability, namely :

- Additive law of probability or theorem of total probability.

## Theory of Probability

b) Multiplicative law of probability or theorem of compound probability.

**Additive Law of probability (for mutually exclusive events) :** The probability of one of the several mutually exclusive events  $A_1, A_2, \dots, A_k$  will happen is the sum of the individual probabilities of the separate events. Symbolically

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

**Proof.** Let  $n$  be the total number of exhaustive, equally likely and mutually exclusive cases of which  $m_1, m_2, \dots, m_k$  are respectively the favourable number of cases to the events  $A_1, A_2, \dots, A_k$ . Since the events are disjoint i. e. non-overlapping then the total number of cases favourable to the events either  $A_1$  or  $A_2$  or ..... or  $A_k$  is  $m_1 + m_2 + \dots + m_k$ .

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

$$= \frac{m_1}{n} + \frac{m_2}{n} + \dots + \frac{m_k}{n} = P(A_1) + P(A_2) + \dots + P(A_k) \quad \dots \dots \dots (5.1)$$

Hence the theorem is proved.

In set notation, the law can be written as

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k) \quad \dots \dots \dots (5.2)$$

**Additive Law of Probability (for not mutually exclusive events) :** The probability of one of the several not mutually exclusive events  $A_1, A_2, \dots, A_k$  is given by

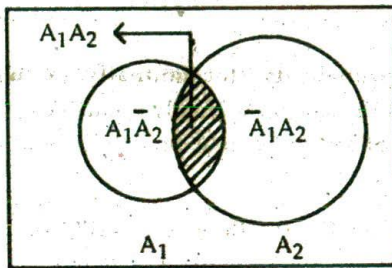
$$P(A_1 + A_2 + \dots + A_k) = \sum_{i=1}^k P(A_i) - \sum_{\substack{i,j=1 \\ i \neq j}}^k P(A_i A_j) + \dots + (-1)^{k-1} P(A_1 A_2 \dots A_k).$$

**Proof :** Let us consider two not mutually exclusive events  $A_1$  and  $A_2$  initially. Since the event  $A_1$  can occur either by two exhaustive mutually exclusive forms  $A_1 A_2$  and  $A_1 \overline{A_2}$ , where  $\overline{A_2}$  indicates not happening of  $A_2$ . Then

$$P(A_1) = P(A_1 A_2) + P(A_1 \overline{A_2}) \quad \dots \dots \dots (5.3)$$

Similarly  $P(A_2) = P(A_1 A_2) + P(\overline{A_1} A_2) \quad \dots \dots \dots (5.4)$

An Introduction to The Theory of Statistics



Adding (5.3) and (5.4) we get

$$\begin{aligned}
 P(A_1) + P(A_2) &= P(A_1A_2) + P(A_1\bar{A}_2) + P(\bar{A}_1A_2) + P(\bar{A}_1\bar{A}_2) \\
 &= P(A_1A_2) + P(A_1 + A_2) \\
 \therefore P(A_1 + A_2) &= P(A_1) + P(A_2) - P(A_1A_2) \quad \dots\dots\dots(5.5)
 \end{aligned}$$

In set notation (5.5) can be written as

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \quad \dots\dots\dots(5.6)$$

Again let us consider two events \$A\_1\$ and \$(A\_2 + A\_3)\$ we have by (5.5)

$$\begin{aligned}
 P(A_1 + (A_2 + A_3)) &= P(A_1) + P(A_2 + A_3) - P\{A_1(A_2 + A_3)\} \\
 &= P(A_1) + P(A_2) + P(A_3) - P(A_2A_3) - P\{A_1A_2 + A_1A_3\} \\
 &= P(A_1) + P(A_2) + P(A_3) - P(A_2A_3) - P(A_1A_2) - P(A_1A_3) + P(A_1A_2A_3).
 \end{aligned}$$

Since the event \$A\_1A\_2A\_1A\_3 = A\_1A\_2A\_3\$ then \$P(A\_1A\_2A\_1A\_3) = P(A\_1A\_2A\_3)\$.

Therefore, \$P(A\_1 + A\_2 + A\_3) = P(A\_1) + P(A\_2) + P(A\_3) - (A\_2A\_3) - P(A\_1A\_3) - P(A\_1A\_2) + P(A\_1A\_2A\_3)\$.

.....(5.7)

The above result can be generalised as

$$P(A_1 + A_2 + \dots + A_k) = \sum_{i=1}^k P(A_i) - \sum_{\substack{i,j=1 \\ i \neq j}}^k P(A_iA_j) + \dots + (-1)^{k-1} P(A_1A_2\dots A_k) \quad \dots\dots(5.8)$$

In set notation we have,

$$\begin{aligned}
 P(A_1 \cup A_2 \cup \dots \cup A_k) &= \sum_{i=1}^k P(A_i) - \sum_{\substack{i,j=1 \\ i \neq j}}^k P(A_i \cap A_j) \\
 &\quad \dots\dots + (-1)^{k-1} P(A_1 \cap A_2 \dots \cap A_k) \quad \dots\dots(5.9)
 \end{aligned}$$

**Example 5.2** An urn contains 3 red, 3 black and 6 white identical balls. Three balls are drawn at random. What is the probability that all the balls are of same colour?



## Theory of Probability

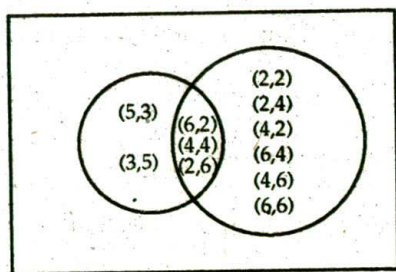
**Solution :** Set  $A_1$  be the event that three balls will be red,  $A_2$  be the event that three balls will be black and  $A_3$  be the event that three balls will be white. Here the events are mutually exclusive, and we know  $P(A_1 + A_2 + A_3) = P(A_1) + P(A_2) + P(A_3)$

$$\text{Here } P(A_1) = \frac{{}^3C_3}{{}^{12}C_3} = \frac{1}{220} \quad P(A_2) = \frac{{}^3C_3}{{}^{12}C_3} = \frac{1}{220} \quad \text{and } P(A_3) = \frac{{}^6C_3}{{}^{12}C_3} = \frac{20}{220}$$

$$\text{Hence the required probability is } \frac{1}{220} + \frac{1}{220} + \frac{20}{220} = \frac{22}{220} = \frac{1}{10}$$

**Example 5.3** Two unbiased dice are tossed simultaneously. What is the probability of getting a total of point 8 or even numbers from both the dice?

**Solution :** Let  $A_1$  be the event of getting a total of point 8 and  $A_2$  be the event of getting even numbers from both the dice. Here the events  $A_1$  and  $A_2$  are not mutually exclusive because a total of 8 points can be had from some of the even numbers from both the dice and is shown in the following Venn diagram.



There are  $6 \times 6 = 36$  points in the universal set when 2 dice are tossed.

$$\text{We have, } P(A_1) = \frac{5}{36}, \quad P(A_2) = \frac{9}{36} \quad \text{and } P(A_1 \cap A_2) = \frac{3}{36}$$

$$\text{Hence, } P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = \frac{5}{36} + \frac{9}{36} - \frac{3}{36} = \frac{11}{36}$$

Therefore the required probability is  $\frac{11}{36}$ .

**Compound Event :** Two events A and B are said to be compound event AB if they are connected and may occur simultaneously. Similarly the compound event ABC.....etc. can be defined. And the probability of the compound event AB is denoted by  $P(AB)$ . similarly  $P(ABC)$  etc., can also be defined.

**Marginal Probability and Conditional Probability :** Let us consider an experiment which can give result in the occurrence of events A and  $\overline{A}$  and also result in the occurrence of B and  $\overline{B}$ . The number of cases favourable to the event AB,  $\overline{A}B$ ,  $A\overline{B}$  and  $\overline{A}\overline{B}$  are  $n_{11}$ ,  $n_{21}$ ,  $n_{12}$  and  $n_{22}$  respectively. And  $n=(n_{11}+n_{21}+n_{12}+n_{22})$  be the total number of cases are shown as in Table -5.2

Table-5.2

	A	$\overline{A}$	Total
B	$n_{11}$	$n_{21}$	$n_{.1}$
$\overline{B}$	$n_{12}$	$n_{22}$	$n_{.2}$
Total	$n_{.1}$	$n_{.2}$	$n$

In the Table 5.2  $n_{.1}$ ,  $n_{.2}$ ,  $n_{.1}$  and  $n_{.2}$  are called the marginal totals.

The probability of happening of the event A, denoted by  $P(A) = \frac{n_{.1}}{n}$

Similarly

$$P(\overline{A}) = \frac{n_{.2}}{n} \quad P(B) = \frac{n_{.1}}{n} \quad \text{and} \quad P(\overline{B}) = \frac{n_{.2}}{n}$$

These probabilities are called marginal probabilities.

Total number of cases for the event B is  $n_{.1}$  and the number of favourable cases to the event A when B has already occurred is  $n_{11}$ . Therefore, the probability of A given that B has already occurred is  $\frac{n_{11}}{n_{.1}}$ . This probability is called the conditional probability of A given that B has already occurred and is denoted by  $P(A/B)$ . Similarly other conditional probabilities from Table 5.2 can be defined.

**Independent and Dependent Events :** Events are said to be independent if the happening or non-happening of an event is not affected by the happening of any number of remaining events. Otherwise the events are said to be dependent.

For example, in case of drawing of a card from a well shuffled pack of cards and replaces it before drawing the second card, the result of the second draw

## Theory of Probability

is independent of the first draw. If the first card drawn is not replaced then the second draw is dependent on the first draw.

Two events A and B are said to be independent if any one of the following conditions is satisfied.

- i)  $P(AB) = P(A) P(B)$ .
- ii)  $P(A/B) = P(A)$ .
- iii)  $P(B/A) = P(B)$ .

**Multiplicative Law of Probability (for dependent events) :** The probability of the simultaneous occurrence of two dependent events A and B is equal to the probability of A multiplied by the conditional probability of B given that A has already occurred (or it is equal to the probability of B multiplied by the conditional probability of A given that B has already occurred). Symbolically,  $P(AB) = P(A) P(B/A) = P(B) P(A/B)$ .

**Proof :** Let n denote the total number of mutually exclusive and equally likely cases of which  $n_1$  cases are favourable to the event A. The cases favourable to both A and B is  $n_{11}$  which is included in  $n_1$ . Then the probability of happening both the events A and B, denoted by  $P(AB)$  is given by

$$P(AB) = \frac{n_{11}}{n} = \frac{n_1}{n} \times \frac{n_{11}}{n_1}$$

The ratio  $\frac{n_1}{n}$  is the probability of the event A, denoted by  $P(A)$  and the ratio  $\frac{n_{11}}{n_1}$  is the conditional probability of B given that A has already occurred, denoted by  $P(B/A)$ .

Hence  $P(AB) = P(A) P(B/A)$ . .....(5.10)

In the compound event AB, the order of the letters are immaterial.

Hence we may write,  $P(AB) = P(B) P(A/B)$ .

**Remarks :**

- 1)  $P(A/B) = \frac{P(AB)}{P(B)}$  if  $P(B) > 0$
- 2) For three events A, B and C,  
 $P(ABC) = P(A) P(B/A) P(C/AB)$ . .....(5.11)

Thus the theorem can be generalised.

- 3) If A and B are independent events then from the condition of independence we know,  $P(B/A) = P(B)$ .



## An Introduction to The Theory of Statistics

Hence  $P(AB) = P(A) P(B)$ . .....(5.12)

For three independent events A, B and C,

$P(ABC) = P(A).P(B).P(C)$ . ....(5.13)

Thus the theorem can be generalised:

**Example 5.4** A bag contains 4 white and 5 red balls. Two balls are drawn successively at random from the bag. What is the probability that both the balls are white when the drawings are made,

- i) with replacement,    ii) without replacement.

**Solution :** Let A be the event that the first ball is white and B be the event that the second ball is also white.

- i) Since the drawings are made with replacement, the two events become independent.

$$\text{Hence } P(AB) = P(A) P(B) = \frac{4}{9} \cdot \frac{4}{9} = \frac{16}{81}$$

- ii) Since the drawings are made without replacement the events become dependent.

$$\text{Hence } P(AB) = P(A) P(B/A) = \frac{4}{9} \cdot \frac{3}{8} = \frac{12}{72} = \frac{1}{6}$$

**Example 5.5** Three groups of children contains respectively 3 girls and 1 boy, 2 girls and 2 boys, 1 girl and 3 boys. One child is selected at random from each group. Find the probability that the three selected children consists of 1 girl and 2 boys.

**Solution :** Let G represent girl, B represent boy, also let the sequence GBB indicate the group of children having girl from first group, boy from second and third groups respectively, similarly BGB and BBG can also be defined.  $P(GBB)$  indicates the probability of selecting a group with one girl from first group and 2 boys each from second and third groups. The event of selecting a group consisting 1 girls and 2 boys may happen either by GBB or BGB or BBG.

Since the event of selecting one girl from first group is independent of the event of selecting one boy from second group and also the event of selecting one boy from third group, we have,  $P(GBB) = P(G) P(B) P(B)$ .

### Theory of Probability

Again since the groups GBB, BCB and BBC are mutually exclusive, the probability of selecting a group consisting 1 girl and 2 boys is the sum of the probabilities of the above three groups.

$$\text{Now, } P(\text{GBB}) = \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{18}{64}, \quad P(\text{BCB}) = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{6}{64} \quad \text{and} \quad P(\text{BBC}) = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{2}{64}.$$

$$\text{Hence the required probability is } \frac{18}{64} + \frac{6}{64} + \frac{2}{64} = \frac{26}{64} = \frac{13}{32}.$$

**Bayes' Theorem :** If  $B_1, B_2, \dots, B_n$  are  $n$  mutually exclusive events with  $P(B_i) \neq 0, (i = 1, 2, \dots, n)$  then for an event  $A$  that occurs when the experiment is performed, such that  $P(A) > 0$ , we have

$$P(B_i/A) = \frac{P(B_i) P(A/B_i)}{\sum_{i=1}^n P(B_i) P(A/B_i)} \quad \dots \dots \dots (5.14)$$

**Proof :** Let us suppose that the probabilities  $P(B_1), P(B_2) \dots P(B_n)$  and the conditional probability  $P(A/B_i)$  are known.

By the theorem of compound probability, we have,

$$P(AB_i) = P(A | B_i) \cdot P(B_i)$$

$$\therefore P(B_i/A) = \frac{P(B_i) \cdot P(A/B_i)}{P(A)} \quad \dots \dots \dots (5.15)$$

Here the event  $A$  can happen in any of the mutually exclusive cases  $AB_1, AB_2, \dots, AB_n$ . By the theorem of total probability we have,

$$\begin{aligned} P(A) &= P(AB_1) + P(AB_2) + \dots + P(AB_n) \\ &= P(B_1) P(A/B_1) + P(B_2) P(A/B_2) + \dots + P(B_n) P(A/B_n) \\ &= \sum_{i=1}^n P(B_i) P(A/B_i) \quad \dots \dots \dots (5.16) \end{aligned}$$

Therefore, from (5.15) and (5.16) we have

$$P(B_i/A) = \frac{P(B_i) P(A/B_i)}{\sum_{i=1}^n P(B_i) P(A/B_i)} \quad \text{Hence proved.}$$

**Remark :**

- 1)  $P(B_i)$ , ( $i=1, 2, \dots, n$ ) are known as a-priori probabilities.
- 2)  $P(B_i/A)$  is called a-posteriori probability.

**Example 5.6** There are two identical boxes containing respectively 4 white and 3 red balls, 3 white and 7 red balls. A box is chosen at random and a ball is drawn at random from it. If the ball is white, what is the probability that it is from the first box?

**Solution :** Let  $B_1$  be the event that the first box is chosen and  $B_2$  be the event that the second box is chosen and  $A$  be the event of getting a white ball. As the boxes are identical and are chosen at random,

$$P(B_1) = P(B_2) = \frac{1}{2}; \quad P(A/B_1) = \frac{4}{7} \quad \text{and} \quad P(A/B_2) = \frac{3}{10}.$$

Hence the required probability,

$$P(B_1/A) = \frac{P(B_1) P(A/B_1)}{P(B_1) P(A/B_1) + P(B_2) P(A/B_2)}$$

$$= \frac{\frac{1}{2} \times \frac{4}{7}}{\frac{1}{2} \times \frac{4}{7} + \frac{1}{2} \times \frac{3}{10}}$$

$$= \frac{40}{61}$$



## 6. RANDOM VARIABLES AND PROBABILITY DISTRIBUTION FUNCTIONS

### 6.1 Random Variables

We have discussed the concept of variables in chapter 2. A random variable or simply a variable must have a range or set of possible values associated with a definite probability with each value.

Let us consider an example of four possible points obtained by tossing 2 unbiased coins simultaneously as shown below :

HH, HT, TH, TT where H indicates that the coin shows head upward and T indicates tail upward. Here the number of heads,  $x$ , are 0, 1 and 2 with corresponding probabilities  $\frac{1}{4}$ ,  $\frac{1}{2}$  and  $\frac{1}{4}$  which constitute a random variable.

Random variable are usually denoted by capital letters e. g. X, Y, Z.....etc. and the values of these variables are denoted by small letters x, y, z .... etc. but in this text we use small letters to represent random variables and their values as well.

A random variable is called discrete if it assumes only a finite number and a random variable is continuous if it takes all possible values between certain limits.

### 6.2 Probability Functions

The probability  $p(x_i)$  associated with values of discrete random variable  $x_i$  is called probability function. Thus the probability function of the above example is

$$P(x_i) = {}^2C_{x_i} \left(\frac{1}{2}\right)^2; \quad i = 0, 1, 2.$$

The probability function is usually denoted by a formula rather than giving some numerical values. A probability function  $P(x_i)$  should satisfy the following conditions :

- i)  $P(x_i) \geq 0$ , for all admissible  $i$

ii)  $\sum_{i=1}^n P(x_i) = 1$ , where the random variable assumes  $n$  values

like  $x_1, x_2, \dots, x_n$ .

### 6.3 Probability Density Function

For continuous random variable  $x$ , the probability that it will be within the small interval  $\left[ x - \frac{dx}{2}, x + \frac{dx}{2} \right]$  of length  $dx$  round the point of  $x$  and is denoted by  $f(x)dx$ . The function  $f(x)$  is usually known as probability density function (p. d. f) which satisfies the following conditions :

i)  $f(x) \geq 0$  for all  $x$  within the range,

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

iii) The probability that the continuous variable  $x$  with p. d. f  $f(x)$  fall in any interval  $(a, b)$  is given by

$$\text{Prob } (a \leq x \leq b) = \int_a^b f(x) dx.$$

**Example 6.1** A random variable  $x$  has the following probability density function,  $f(x)=cx(2-x), 0 \leq x \leq 2$

i) determine  $c$ , ii) find the probability that  $0 \leq x \leq 1$ .

**Solution :** i) Since we know  $\int_0^2 f(x) dx=1$ , the value of  $c$  can be obtained

as,  $c \int_0^2 (2x-x^2) dx=1$ .

or,  $c \left[ \frac{2x^2}{2} - \frac{x^3}{3} \right]_0^2 = 1$  or,  $c \left[ 4 - \frac{8}{3} \right] = 1$

or,  $c \frac{4}{3} = 1 \therefore c = \frac{3}{4}$

$$\text{ii) Prob. } (0 \leq x \leq 1) = \frac{3}{4} \int_0^1 (2x-x^2) dx$$

$$= \frac{3}{4} \left[ \frac{2x^2}{2} - \frac{x^3}{3} \right]_0^1 = \frac{3}{4} \left[ 1 - \frac{1}{3} \right] = \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2}$$

#### 6.4 Different Measures of Central Location for Continuous Probability Distribution

Let  $f(x)$  be the probability density function, (p.d.f) of a random variable  $x$ ,  $a \leq x \leq b$  then

$$1) \text{ Arithmetic Mean} = \int_a^b x f(x) dx$$

$$2) \text{ Harmonic Mean (H)} : \frac{1}{H} = \int_a^b \frac{1}{x} f(x) dx.$$

$$3) \text{ Geometric Mean (G)} : \log G = \int_a^b \log x f(x) dx.$$

4) Median ( $M_c$ ) : Since median ( $M_c$ ) divides the entire distribution into two equal parts then

$$\int_a^{M_c} f(x) dx = \int_{M_c}^b f(x) dx = \frac{1}{2}$$

Thus solving any one of the above two integrals we get the value of  $M_c$ .

5) Mode ( $M_o$ ) : Mode is that value of  $x$  for which  $f(x)$  is maximum. Thus  $M_o$  is given by

$$f'(x) = 0$$

$$f''(x) < 0$$

Provided that the solution of  $f'(x) = 0$  lies within the permissible range of the variable.



6. Moments: The  $r$ th raw moment,  $\mu'_r = \int_a^b x^r f(x) dx$ ,

and the  $r$ th corrected moment,  $\mu_r = \int_a^b (x - \mu)^r f(x) dx$ , where  $\mu$  is the arithmetic mean of the distribution.

**Example 6-2** Find harmonic mean, mode and median from the probability density function given in Example 6.1.

**Solution:** We know that p. d. f,  $f(x) = \frac{3}{4}x(2-x)$ .

Harmonic mean  $H$  is given by,

$$\begin{aligned} \frac{1}{H} &= \frac{3}{4} \int_0^2 \frac{1}{x} \cdot x(2-x) dx = \frac{3}{4} \int_0^2 (2-x) dx \\ &= \frac{3}{4} \left[ 2x - \frac{x^2}{2} \right]_0^2 = \frac{3}{4} [4-2] = \frac{3}{2} \quad \therefore H = \frac{2}{3} \end{aligned}$$

If  $M_e$  is the median then,  $\frac{3}{4} \int_0^{M_e} x(2-x) dx = \frac{1}{2}$

$$\text{or, } \frac{3}{4} \int_0^{M_e} (2x - x^2) dx = \frac{1}{2}$$

$$\text{or, } \frac{3}{4} \left[ \frac{2x^2}{2} - \frac{x^3}{3} \right]_0^{M_e} = \frac{1}{2}$$

$$\text{or, } \left( M_e^2 - \frac{M_e^3}{3} \right) = \frac{2}{3}$$

$$\text{or, } M_e^3 - 3M_e^2 + 2 = 0$$

$$\text{or, } (M_e - 1)(M_e^2 - 2M_e - 2) = 0.$$

The only value of  $M_e$  lying within the range  $[0, 2]$  is  $M_e = 1$ .

Hence median = 1.

Mode: We know,  $f'(x) = 0$

## Random Variables and Probability Distribution Functions

$f''(x) < 0$  will give the solution for mode within the admissible range, hence,

$$f'(x) = 2 - 2x = 0 \quad \text{or, } 2(1 - x) = 0$$

$$\text{or, } x=1$$

and  $f''(x) = -2$ , Therefore, mode = 1.

### 6.5 Distribution Function

Let  $x$  be a random variable, discrete or continuous and  $F(x)$  be the probability that the random variable  $x$  takes values less than or equal to  $x$  then  $F(x)$  is called the cumulative distribution function (c. d. f) or simply distribution function (d. f) of  $x$ .

Probability function and distribution function of the random variable  $x_i = 0, 1, 2$  obtained by tossing 2 unbiased coins simultaneously can be shown in Table -6.1.

Table -6.1

$x_i$	0	1	2
$P(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$F(x_i)$	$\frac{1}{4}$	$\frac{3}{4}$	1

For discrete random variable  $x_i$  having values  $x_1, x_2, \dots, x_n$ ,

$F(x_k) = \text{Prob} \{x_i \leq x_k\} = P(x_1) + P(x_2) + \dots + P(x_k)$ , ( $k < n$ ). and for continuous random variable  $x$ ,  $-\infty \leq x \leq \infty$ ,

$$F(x_k) = \text{Prob} \{x \leq x_k\} = \int_{-\infty}^{x_k} f(x) dx.$$

**Properties of Distribution Functions :** Following are the properties of distribution function :

- $F'(x) = f(x) \geq 0$ , so that  $F(x)$  is a non-decreasing function.
- $F(-\infty) = 0$
- $F(\infty) = 1$
- $\text{Prob} \{a \leq x \leq b\} = F(b) - F(a)$ .

**Remark :** Properties (ii) and (iii) imply that  $0 \leq F(x) \leq 1$ .

### 6.6 Joint Probability Function

Discrete random variables  $x, y$  etc. are said to be jointly distributed if they are defined on a same probability space. In particular, if we consider only two discrete random variables  $x$  and  $y$ , we have a joint probability function  $P(x_i, y_j)$ , called bivariate probability function which follows the following conditions :

- i)  $P(x_i, y_j) \geq 0$ , for all admissible  $i$  and  $j$ .
- ii)  $\sum_i \sum_j P(x_i, y_j) = 1$ .

**Marginal Probability Function :** Let  $P(x_i, y_j)$  be the probability function of the discrete random variables  $x$  and  $y$ , then the marginal probability function of  $x, P(x_i)$  is given by

$$P(x_i) = \sum_j P(x_i, y_j), \text{ for all } i.$$

Similarly the marginal probability function of  $y, P(y_j)$  is given by

$$P(y_j) = \sum_i P(x_i, y_j), \text{ for all } j.$$

**Conditional Probability Function :** The conditional probability function of the discrete random variable  $x$  for a given value of  $y, P(x_i/y_j)$  is given by

$$P(x_i/y_j) = \frac{P(x_i, y_j)}{P(y_j)}.$$

Similarly conditional probability of  $y$  for given  $x,$

$$P(y_j/x_i) = \frac{P(x_i, y_j)}{P(x_i)}.$$

**Remark :** Two discrete random variates  $x$  and  $y$  are said to be independent if any one of the following conditions is satisfied :

- a)  $P(x_i, y_j) = P(x_i) P(y_j)$
- b)  $P(x_i/y_j) = P(x_i)$
- c)  $P(y_j/x_i) = P(y_j)$ .

**Example 6.3** The joint probability function of  $x$  and  $y$  is

$$P(x, y) = c(x+y); \quad x = 1, 2, 3, \quad y = 1, 2$$

Find i) the value of  $c,$

ii) the marginal probability function of  $x$  and  $y,$



## Random Variables and Probability Distribution Functions

- iii) conditional probability function of  $x$  for given  $y$  and that of  $y$  for given  $x$ .

**Solution :** i) We know,  $\sum_{x=1}^3 \sum_{y=1}^2 c(x+y) = 1$

or,  $c(1+1+1+2+1+3+2+1+2+2+2+3) = 1$ .

or,  $21c = 1 \quad \therefore c = \frac{1}{21}$ .

Hence,  $P(x,y) = \frac{(x+y)}{21}$ .

- ii) The marginal probability functions are

$$P(x) = \sum_{y=1}^2 \frac{(x+y)}{21} = \frac{x+1+x+2}{21} = \frac{2x+3}{21}; \quad x=1, 2, 3$$

and  $P(y) = \sum_{x=1}^3 \frac{(x+y)}{21} = \frac{y+1+y+2+y+3}{21} = \frac{3y+6}{21} = \frac{y+2}{7}; \quad y=1, 2$ .

- iii) The conditional probability functions are

$$P(x/y) = \frac{P(x,y)}{P(y)} = \frac{\frac{x+y}{21}}{\frac{y+2}{7}} = \frac{x+y}{3(y+2)}$$

and  $P(y/x) = \frac{P(x,y)}{P(x)} = \frac{\frac{x+y}{21}}{\frac{2x+3}{21}} = \frac{x+y}{2x+3}$

### 6.7 Joint Probability Density Function

For continuous random variables  $x, y, \dots$  etc. the joint probability density function usually denoted by  $f(x, y, \dots)$  which is the probability that they will be within the small interval  $dx, dy, \dots$  etc. round the point  $x, y, \dots$  etc. In particular if we consider two variables  $x$  and  $y$ , the joint probability density function  $f(x,y) dx dy$  represent the probability that a random point  $(x', y')$  will fall within an infinitesimal region such that  $x < x' < x + dx$  and  $y < y' < y + dy$ .

The function  $f(x,y)$  is called bivariate probability density function.

The function  $f(x,y)$  must follow the following conditions :

- i)  $f(x,y) \geq 0$

$$\text{ii) } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1.$$

**Marginal Probability Density Function :** The marginal probability density function of  $x$  and  $y$  are

$$f(x) = \int_{-\infty}^{\infty} f(x,y) dy ; -\infty \leq x \leq \infty.$$

$$\text{and } f(y) = \int_{-\infty}^{\infty} f(x,y) dx ; -\infty \leq y \leq \infty.$$

**Conditional Probability Density Function :** The conditional probability density function of  $x$  for given  $y$  which is denoted by  $f(x/y)$  and that of  $y$  for given  $x$ , denoted by  $f(y/x)$  are

$$f(x/y) = \frac{f(x,y)}{f(y)}$$

$$\text{and } f(y/x) = \frac{f(x,y)}{f(x)}$$

**Remark :** Two continuous random variables  $x, y$  are said to be independent if any one of the following conditions is satisfied.

i)  $f(x,y) = f(x) f(y).$

ii)  $f(x/y) = f(x).$

iii)  $f(y/x) = f(y).$

**Example 6.4** If the joint p. d. f.  $f(x,y) = xe^{-x(y+1)}$ ;  $x \geq 0, y \geq 0$ ; find marginal and conditional density functions and also show that  $x$  and  $y$  are dependent.

**Solution :** Marginal p. d. f of  $x$  is given by

$$\begin{aligned} f(x) &= \int_0^{\infty} x e^{-x(y+1)} dy \\ &= -x \left[ \frac{e^{-x(y+1)}}{-x} \right]_0^{\infty} \\ &= e^{-x(y+1)} \Big|_0^{\infty} = e^{-x}; x \geq 0. \end{aligned}$$

Marginal p. d. f of  $y$  is given by

## Random Variables and Probability Distribution Functions

$$f(y) = \int_0^{\infty} x e^{-x(y+1)} dx$$

Putting  $x(y+1) = z$ , or  $x = \frac{z}{y+1}$  and  $dx = \frac{dz}{y+1}$ ,  $z \geq 0$

$$\begin{aligned} &= \int_0^{\infty} \frac{z}{(y+1)} \frac{dz}{(y+1)} e^{-z} = \frac{1}{(y+1)^2} \int_0^{\infty} e^{-z} z dz \\ &= \frac{1}{(y+1)^2} \left[ 2 = \frac{1}{(y+1)^2} \right]; y \geq 0. \text{ since } \int_0^{\infty} z e^{-z} dz = 1. \end{aligned}$$

Conditional p. d. f of  $x$  for given  $y$ ,  $f(x/y)$  is given by

$$f(x/y) = \frac{x e^{-x(y+1)}}{1} = (1+y)^2 x e^{-x(y+1)}; x, y \geq 0.$$

Conditional p. d. f of  $y$  for given  $x$ ,  $f(y/x)$  is given by

$$f(y/x) = \frac{x e^{-x(y+1)}}{e^{-x}} = x e^{-xy}; \quad x, y \geq 0.$$

Since  $f(x) \neq f(x/y)$  or  $f(y) \neq f(y/x)$ ,  $x$  and  $y$  are dependent variate.

**Example 6.5** The joint probability density function of two random variables  $x$  and  $y$  is given by

$$f(x, y) = 4xy; \quad 0 \leq x, y \leq 1.$$

Find  $f(x)$ ,  $f(y)$ ,  $f(x/y)$ ,  $f(y/x)$  and check whether  $x$  and  $y$  are independent variables.

**Solution:** We know,

$$f(x) = \int_0^1 4xy dy = 4x \left[ \frac{y^2}{2} \right]_0^1 = 2x.$$

$$f(y) = \int_0^1 4xy dx = 4y \left[ \frac{x^2}{2} \right]_0^1 = 2y.$$

$$f(x/y) = \frac{4xy}{2y} = 2x \text{ and } f(y/x) = \frac{4xy}{2x} = 2y$$

Two variates  $x$  and  $y$  are independent as  $f(xy) = f(x) \cdot f(y)$ .

and also  $f(x/y) = f(x)$  and  $f(y/x) = f(y)$ .