Chapter II : 12

CHAPTER I

INTRODUCTION AND INTRODUCTORY CONCEPTS

Statistics is a numerical description of some events or subjects. Secondly, it is a method of analysis and interpretation of data. However, statistics usually refers to techniques and methods. That is, statistics is a branch of knowledge which includes appropriate method of collection of data on certain problem, its presentation and analysis, and finding out the truth from the results of the analysis. Universal and complete definition of statistics is not available. Users have defined statistics from their own application view point. Definitions given by three famous scientists are stated below :

According to **Prof. A.L. Bowley**, "Statistics is a science of measurement of social organism regarded as a whole in all its manifestation".

Webster defined statistics as "the classified facts relating to the condition of the people in a state especially those facts which can be stated in numbers or in table of numbers or in any tabular or classified arrangement".

Webster's definition is limited only in the information on the general conditions of the people of a state; it does not include information on the other branches of knowledge. But modern statistics has included all spheres of human activities. According to Bowley's definition, statistical science refers to mutually distributed numerical data in any field of investigation. This definition does not cover the causes that affect the correctness of the information in the data collection process.

However, **Prof. H. Secrist** has given a definition of statistics covering all its application fields. According to him "statistics means the aggregate of facts affected to a considerable extent by multiplicity

of causes, numerically expressed, enumerated or estimated according to reasonable levels of accuracy, collected in a systematic method for a predetermined purpose and placed in relation to each other". This can be considered as a complete definition of statistics.

The different functions of statistics are -

- Collection of data
- Organization of data
- Presentation of data
- Analysis of data
- Interpretation and drawing inference.

Thus "**statistics can also be defined as a branch of science which deals with collection, organization, presentation and analysis of numerical data and drawing valid inference therefrom**".

Characteristic Features of Statistics :

In order to help the reader in understanding the nature of the subject some of the characteristic features of the science of statistics are stated below :

i. Statistics deals with aggregate of individuals rather than with individuals. Per capita income of a country is a statistical information because it is an information about the population.

ii. Statistics deals with variation.

iii. Statistics deals with only numerically specified populations.

iv. Statistical inferences are drawn with the probability or uncertainty.

v. The logic used in statistical inference is inductive.

1.1. Uses of Statistics :

Statistics is such a science, application of which is inevitable in all spheres of life. Statistics has wide application in solution of

problems related to Economics, Social Science, Biological Science, Agricultural Science, Business, Planning, Education and Research. There is no branch of knowledge whose data analysis does not require statistical techniques. Application of statistics in some important fields are briefly discussed in this section.

(i) Statistics in Agriculture :

Statistics is widely used in the field of agriculture. In any country, particularly in an agro-based country, development planning and its implementation is very important. Agricultural census are conducted to collect different information on agriculture. Data on total cultivable land, cultivated and cultivable crops, fertilizer use, irrigation etc. are collected, organized and presented. Statistical techniques are applied in analysing these data and on the basis of such analysis regionwise requirements of fertilizer, irrigation etc. for different crops are estimated. Besides, statistical techniques are essential for collection and analysis of data on domestic animals and birds, in respect of their numbers, food, health, growth etc. Various information are collected and analysed even on wild animals using statistical techniques and analysis of various data on veterinary treatment, fish culture, etc. Above all, statistics plays a very important role in the overall agricultural development.

(ii) Statistics in Economics :

Statistics has an intensive relationship with Economics. Statistical techniques are widely used in formulation of economic rules and in testing their effectiveness. In fact, statistical methods has made economics more intensified and popularised. Through application of statistical techniques in the field of Mathematical Economics, a new and popular branch of Economics, named Ecomometrics has been created.

(iii) Statistics in Planning :

Developing countries essentially designs development plans for overall national development. The present age, is often termed as the age of planning. Use of statistics is inevitable in designing such development programmes. National policies are prepared through collection and analysis of various data on peoples' standard of living of country, poverty level, education, trade, different management etc. using statistical methods.

(iv) Statistics in Biology :

Statistical techniques are widely used in analysis of various information in different areas of biological sciences; these data relates to births, deaths, breeding, genetics etc. A new branch of statistics named Biometry has been developed to deal with the different biological aspects. Population studies including fertility, mortality, migration etc. involve lot of statistical exercises.

(v) Statistics in Trade and Commerce :

In the present world of competition data collection and analysis are inevitable for progress and development in trade and commerce. Statistical methods are used as essential tools for collection, analysis and interpretation of data on demand & supply of different commodities, trade cycle, consumers taste, principle and purchasing power etc.

1.2. Limitations of Statistics :

In spite of popular uses of statistical methods in different areas of knowledge, there are some limitations of statistics too; those are briefly discussed below :

- Statistics does not refer to the characteristics of an individual, rather it analyses the collected data and refers to the overall results.

- Statistical results are true on the average; for particular case it may not be true. For example suppose in a certain locality 5% crop lands suffers from pest attacks. The crop of a particular land may not be at all infested whereas infestation of crop in some other plots may be 15% or even more.
- Statistics usually collects data through sample survey and comments on the population characteristics on the basis of sample information. Such sample based comments may not be true in some cases.

1.3. Abuses of Statistics :

- Users of statistics must have sufficient expertise on the subject, because improper use of statistical methods may create complicated problems. Untrained and inexperienced users are very likely to take faulty decision.
- Collected data should be analysed through appropriate statistical technique to arrive at correct decision. But favourable decisions are often taken due to ignorance or by purposively manipulation of data.
- Comments made on the basis of insufficient or inadequate data may be faulty.

Therefore, to ensure proper use of statistical methods, the data should be relevant and complete. The users of statistics should also have necessary knowledge and experience on statistical methods.

1.4. Population and Sample :

It is important to distinguish between a population and a sample.

A population (or universe) is defined as the aggregate of the elementary units. Since the number of elementary units is equal to the number of observations, we may say that a population is the aggregate of the observations. For example, assume that 650 students

are admitted to the first semester of the first level at BAU. We are interested on their H.S.C. score. Each student is an elementary unit and the 650 students together comprise the population. Or, we may say that 650 observations (H.S.C. scores of 650 students) is the population. The size of the population is usually denoted by N.

A sample is a set of n observations (elementary units) drawn from the population. This n is known as the size of the sample. As an illustration, if we select 25 students from the population of 650, we have a sample of size 25. Or, we may say that we select 25 observations (H.S.C. scores) from the population of 650 scores (sample and sampling methods will be discussed in chapter IX). Unknown characteristics which refer to populations are called parameters; and that related to samples are called statistic. Populations, as the term is used in statistics, are arbitrarily defined groups. They need not be as large as the one used here as illustration.

1.5. Scales of Measurement :

The theory of measurement consists of a set of separate or distinct theories, each concerning a distinct scales of measurement. The operations admissible on a given set of scores depend on the scale of measurement used. Different scales of measurement are :

1. Nominal or classificatory scale
2. Ordinal or ranking scale
3. Interval scale
4. Ratio scale

Nominal or Classificatory Scale :

Nominal scales are used as measures of identity. Numbers may serve as labels to identify items or classes. When numbers or certain symbols are used to identify the groups to which objects under observation belong, these numbers or symbols constitute the nominal or classificatory scale.

In its simplest form, the numbers carried on the back of athletes represent a nominal scale. Other examples of such scales are the classification of individual into categories. For example, a sample of people under study may be sorted in different categories on the basis of religious belief; (i) Muslim; (ii) Hindu; (iii) Christian; (iv) Buddhists etc. Or, they may be classified on the basis of sex, political party membership, rural-urban and the like. Simple statistics are used with nominal data. For example, the number, proportion or percentage. These classes or categories are mutually exclusive. The only relation involved is that of equivalence. That is, the member of any class or category are equivalent in the property being scaled. The relation is symbolized by the sign "=". Under certain conditions, we can test hypothesis regarding the distribution of cases among categories using the nonparametric statistical test χ^2; the most common measure of association for nominal data is the contingency coefficient, C.

Ordinal or Ranking Scale :

When an ordinal scale is used in measurement, numbers reflect the rank order of the individuals or objects. Objects in one category are not just different from those in the other category of the scale, but there exists some kind of relationship to them. Examples of such relations are higher, more preferred, taller, brighter, smaller, harder, softer etc. Such relations may be designated by the symbol ">" or sometimes by "<". The fundamental difference between a nominal scale and an ordinal scale is that the ordinal scale incorporates not only the relation of equivalence "=" but also the relation greater than ">" or smaller than "<".

Ordinal measures reveal, for instance, which person or object is taller or heavier than the other. But measures do not tell how much taller or how much heavier one is from the other. Statistically not much can be done with ordinal measures except to determine the median and partition values and to compute rank correlation coefficients.

Interval Scale :

The interval scale provides numbers that reflect differences among items. The interval scale has all the characteristics of the ordinal scale, and in addition, provides the distance between any two numbers. That is, we know how large are the intervals (distances) between all objects on the scale. An interval scale is characterised by a common and constant unit of measurement which assigns a real number to all pairs of objects in the ordered set.

Examples on such scales are the centigrade and Fahrenheit thermometers, scores on intelligence tests, etc.

In constructing an interval scale, one must not only be able to specify equivalences, as in a nominal scale, and greater than (or less than) relation, as in ordinal scale, but also be able to specify the ratio of any two intervals. The interval scale is a true quantitative scale. Many statistics are used with interval scales : arithmetic mean, standard deviation and the Pearson's product moment correlation coefficient. Parametric statistical tests such as t -test and F-test are applicable.

Ratio Scale :

The basic difference between this and the preceding type is that ratio scales have an absolute zero. It is true that interval scales (e.g

Fahrenheit and Centigrade) also have zero points, but such points are arbitrarily chosen. Common ratio scales are measures of length, width, weight, capacity, loudness, and so on. When a ratio scale is used numbers reflect ratios among items, and data obtained with such scales may be subjected to the highest type of statistical treatments.

When the data are in terms of meters or centimeters we can say that one length or height is twice or half of that of another. When our measurements are on an interval scale, we cannot do this. For example suppose the maximum temperature today is 60^0; the same day last year it was 30^0. In this case we cannot state that it is twice as warm today as it was on the same date last year. What is the difference between the two conditions ? When where dealing with meter or centimeter, we were using a measuring scale that was based on an absolute zero; in the second case we were using a scale which started 32 degrees below the freezing point of water. When measurements are on the ratio, such meaningful comparisons can be made. As a matter of fact, when data are of this type, all of the mathematical and statistical operations may be made.

Nominal and ordinal measurements are the most common types achieved in behavioral sciences. Data measured by either nominal or ordinal scales should be analysed by the nonparametric methods. Data measured in interval or ratio scales may be analysed by parametric methods, if the assumptions of the parametric statistical model are tenable.

The information in above discussion of different measurement scales and kinds of statistics and statistical tests appropriate to each scale (when the assumptions of the relevant test are met) are summarised below :

Table 1.1 : Scales of Measurement, Appropriate Statistics and Appropriate Statistical Tests

Scale	Defining relation	Examples of appropriate statistics	Appropriate statistical tests
Nominal	(i) Equivalence	Mode Frequency Contingency coefficient	Nonparametric statistical tests
Ordinal	(i) Equivalence (ii) Greater than or less than	Median Quartiles Rank correlation coeff. Kendall's ω	-Do-
Interval	(i) Equivalence (ii) Greater/less than (iii) Known ratio of any two intervals	Mean Standard deviation Simple and multiple correlation coefficients	Nonparametric and parametric statistical tests
Ratio	(i) Equivalence (ii) Greater/less than (iii) Known ratio of any two intervals (iv) Known ratio of any two scale values	Geometric mean Coefficient of variation	-Do-

1.6. Some Statistical Symbols :

We have already mentioned that the unknown characteristics of a population are called parameters, whereas the characteristics of a sample are called statistic. Some usual symbols for parameters and statistics are -

Table 1.2. Population and Sample Characteristics and Their Notations.

Characteristic	Parameter	Statistic
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Variance	σ^2	s^2
Correlation Coefficient	ρ	r
Regression coefficient	β	b

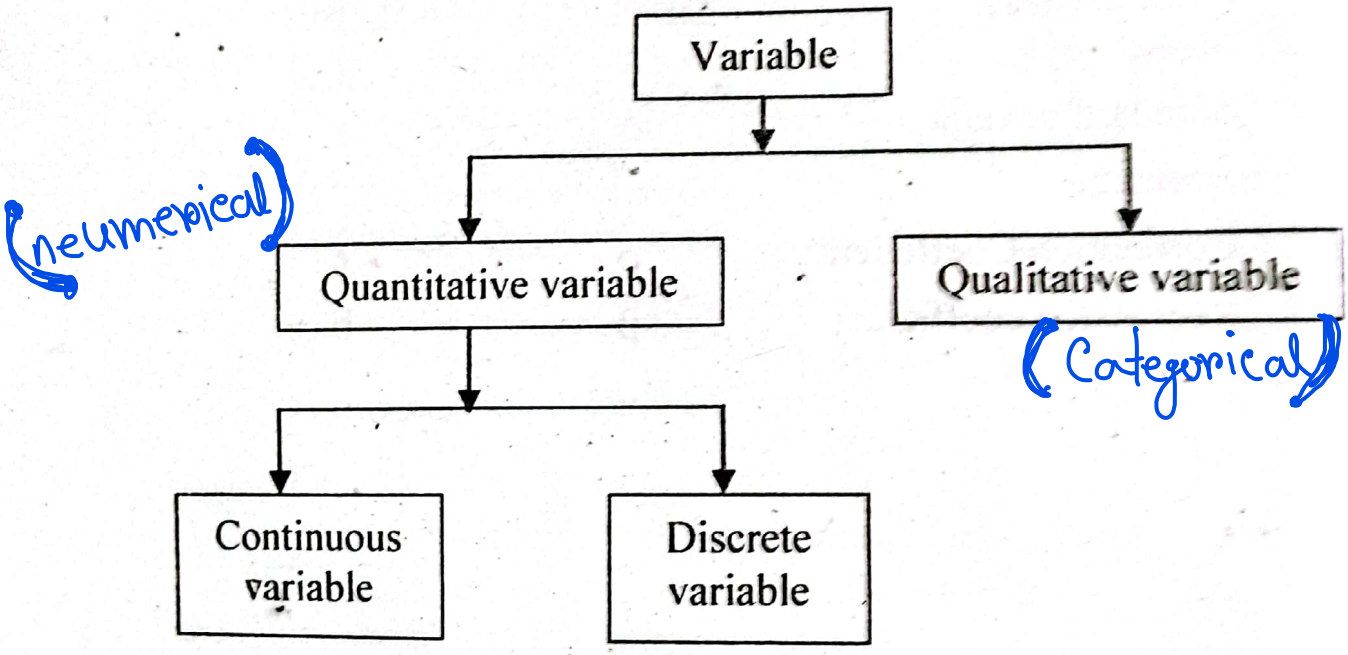
CHAPTER II

VARIABLE AND FREQUENCY DISTRIBUTION

2.1. Variable :

Measurable characteristics of a population that may vary from element to element either in magnitude or in quality are called *variables*. Variables are of two types - *quantitative variable* and *qualitative variable*.

Variables and its classification can be demonstrated as shown below :



Quantitative Variable :

Variable characteristics, whose values are expressed numerically, are known as quantitative variables: Height or weight of students, length or breadth of fishes, weight of tomato, number of grapes per bunch, number of grains per panicle, etc. are some examples of quantitative variables.

Qualitative Variable :

Some variables, which express the quality of population elements, cannot be numerical measured but can be classified

categorised, these are called qualitative variables. For example, merit of students, educational attainment, type of farmers (big, medium, small), type of fishes (sea fish, river fish) etc. cannot be numerically measured but can be grouped into classes or categories. Qualitative variables are also known as attributes.

Quantitative variables are of two types - continuous and discrete.

A variable which can assume any value, integral or fractional, within specified limits, is called a *continuous variable*. For example, height of students, weight of tomato, length of fish, height of trees, weight of animal etc. are continuous variables which can take both integral or fractional values.

On the other hand some variables can take only integral values e.g., number of grains per panicle, number of students per class, number of fishes caught per unit time etc. These are called *discrete variables*.

2.2. Frequency and Frequency Distribution :

In many situations several population elements assume same values; i.e., numerical values of population characteristics may often repeated again and again. For example, several panicles may contain the same number of grains, a number of fishes may have the same length or weight, several students in a class may have the same height. Such repeatation of the value of variable is called frequency. That is, the number of times a particular observation occurs in the data set is the frequency of that particular observation. Frequency is usually denoted by 'f'.

Frequency Distribution :

Information collected in any process are usually classified or grouped according to specific characteristics. Arrangement of

observational data according to frequencies of the observations is called *frequency distribution.*

Frequency distribution should be such that the arrangement according to the observations becomes easily understandable. Frequency distributions are constructed mainly to present the data in condensed form and for easy understanding. Frequency distribution is very important in statistical studies.

Construction of Frequency Distributions :

Steps in constructing a frequency distribution are discussed below :

1. Finding the Range : In constructing frequency distribution the highest and the lowest value in the data set are first identified and their difference is obtained. This difference between the highest value and the lowest value is called the *range* usually denoted by R.

* Range = Highest value - Lowest value. *

2. Decision About the Number of Classes : After finding the range it is necessary to decide the number of classes in which the entire data set should be divided. Choice of the number of classes should be realistic; this number should not be very small and at the same time it should not be very large so that the aim of construction of frequency distribution (condensation) is not achieved. It is generally expected to limit the number of classes between 7 and 15. There is no hard and fast rule for choosing the number of classes. However, M.A. Sturge's formula gives a guideline for desired number of classes. The formula is -

$$k = 1 + 3.322 \log_{10} N$$

where N is the total number of observations in the data set and is the desired number of classes.

3. Choosing the Class Interval : The next step of constructing frequency distribution is the calculation of the class interval. Each class will have two limits, the lower limit (the lower value) and the upper limit (the higher value). The difference of the upper limit and the lower limit of a class is known as *class interval*, usually denoted by c or h. If the range is divided by the number of classes, we get the class interval.

$$\text{Class Interval (C)} = \frac{\text{Range}}{\text{No. of classes}}$$

The value of c is taken as the next integral value of the ratio R/k. For choosing the class interval, there is no rigid rule as to use the exact end values of the data set, rather convenient values near the highest and lowest observations of the data set may be used. However, class interval should be such that classes are distinct and separate from each other. Depending on the nature of the variable, two different methods are used in choosing the class limits. If the variable is discrete, closed intervals like $a \leq x \leq b$ are used, both the lower and upper limits are included (e.g., 0-4, 5-9, 10-14, ... etc.). On the other hand, if the variable is continuous, open interval system ($a < x \leq b$ or $a \leq x < b$) is used; one of the class limits is included and the other is excluded (usually the lower limit is included). In this case the classes will be 0-5, 5-10, 10-15, ..., etc.

4. Counting of Frequencies : For convenience of counting the number of observations falling within each class tally marks are used; frequency of each class is determined by counting the tally marks.

Sometimes it may be necessary to know the observations greater or smaller than a particular value or class of values. For this, cumulative frequencies for observation or class are obtained.

Example 2.1 :

Suppose the marks obtained by 50 students in an examination in Economics are as follows :

32	27	19	40	31	17	15	18	21	27	38	15	33	34	29
26	16	25	33	36	24	22	26	19	36	18	25	20	25	25
31	24	16	28	30	24	29	42	29	28	26	27	47	43	22
25	28	22	24	23										

Here the variable is the marks obtained by the students. The data as shown above are called raw or ungrouped data.

If it is needed to describe the performance of the students, it may be done in a number of ways.

We may enumerate the grade of each student either in ascending or descending order; data such arranged are said to be arranged in array. Counting the number of times each value of the variable occurs, we get a table of the following type :

Table 2.1 Frequency Distribution of Marks

Marks	Frequency (No. of students)	Cumulative frequency	Marks	Frequency (No. of students)	Cumulative frequency
15	2	2	28	3	33
16	2	4	29	3	36
17	1	5	30	1	27
18	2	7	31	2	39
19	2	9	32	1	40
20	1	10	33	2	42
21	1	11	34	1	43
22	3	14	36	2	45
23	2	16	38	1	46
24	4	20	40	1	47
25	5	25	42	1	48
26	3	28	43	1	49
27	2	30	47	1	50

Such a table is known as frequency table or frequency distribution. The above arrangement is an improvement over the raw data, but to get a still better idea of the performance of the students we reclassify the data into grouped frequency distribution as shown below :

Table 2.2: Grouped Frequency Distribution of Marks

Class interval	Tally mark	Frequency (No. of students)	Cumulative frequency	
			Ascending	Descending
15-20	𝍸 ||||	9	9	50
20-25	|||| 𝍸 |	11	20	41
25-30	𝍸 𝍸 ||||	16	36	30
30-35	𝍸 ||	7	43	14
35-40	|||	3	46	7
40-45	|||	3	49	4
45-50	|	1	50	1

This type of classification of raw data is called grouped frequency distribution or simply frequency distribution.

In the above example the highest value is 47, the lowest value is 5 and the range is, $R = 47-15 = 32$.

According to Sturge's formula,

$$k = 1 + 3.322 \log_{10} 50 = 6.47$$

That is, 6 to 7 groups are appropriate in this case.

Again, $C = \frac{R}{k} = \frac{32}{7} = 4.57$; accordingly 5 is taken as the class interval.

Example 2.2: Weight (in gm.) of tomato harvested from the kitch garden are given below :

75	80	52	87	95	105	92	82	120	65
55	100	115	92	82	97	85	72	67	98
115	62	85	98	110	105	77	63	80	90
54	89	108	103	75	53	105	117	95	64
77	85	94	72	68	100	78	89	94	102
82	95	98	100	77	85	92	97	72	85
72	83	66	58	96	75	88	90	80	95
63	78	84	92	88	77	65	85	92	87

In constructing a frequency distribution the highest and lowest observations are to be identified first. In the present data the highest value is 120 and the lowest value is 52. Therefore range

$$R = 120 - 52 = 68 \; ; N = 80$$

According to Sturge's formula

$$k = 1 + 3.322 \log_{10} 80 = 1 + 3.322 \times 1.903089987 = 7.322$$

The next integral value of k is 8; the data set may be groupe about 8 classes.

$$\text{Now, } C = \frac{R}{K} = \frac{68}{8} = 8.5 \approx 9$$

It will be convenient to take 10 as the class interval. As variable here is continuous, open interval method is to be followe grouping the data set. Though the lowest observation is 52, convenient to start from 50. The classes or groups will, therefore 50-60, 60-70, 70-80, 80-90, 90-100, 100-110 and 110-120.

The frequency distribution will be -

Table 2.3: Frequency Distribution of Weight of Tomato

Class interval	Tally mark	Frequency	Cumulative frequency	
			Ascending	Descending
50-60	𝍸	5	5	80
60-70	𝍸 ||||	9	14	75
70-80	𝍸 𝍸 |||	13	27	66
80-90	𝍸 𝍸 𝍸 𝍸	20	47	53
90-100	𝍸 𝍸 𝍸 ||||	19	66	33
100-110	𝍸 ||||	9	75	14
110-above	𝍸	5	80	5

The highest observation 120 appears only once in the data set. This 120 could make another group 120-130. Without introducing a separate group for only one observation, it has been included in the preceding group making it 110-above, instead of 110-120.

Example 2.3 : The following data show the number of grapes per bunch:

25	75	15	20	18	62	45	33	40	45
77	30	35	25	65	42	55	37	44	50
40	35	38	47	52	45	33	28	22	22
18	29	48	55	60	58	43	40	47	39
35	45	43	52	57	50	48	55	59	42
28	36	54	48	58	68	78	61	53	42

Here, N = 60, the highest value is 78 and the lowest value is 15. The variable is discrete in nature.

$R = 78 - 15 = 63$ and $k = 1 + 3.322 \log_{10} 60 = 6.91 \approx 7$.

The data may be classified in more or less 7 groups.

Again, $C = \frac{R}{k} = \frac{63}{7} = 9$

For convenience, however, 10 may be taken as the class interval.

Table 2.4 : Frequency Distribution of No. of Grapes per Bunch

Class interval	Tally mark	Frequency	Cumulative frequency	
			Ascending	Descending
15-24	𝍸 I	6	6	60
25-34	𝍸 III	8	14	54
35-44	𝍸 𝍸 III	15	29	46
45-54	𝍸 𝍸 𝍸 I	16	45	31
55-64	𝍸 𝍸	10	55	15
65-74	II	2	57	5
75-84	III	3	60	3

2.3 Graphical Representation of Frequency Distribution :

Frequency distributions may be presented by graphs and charts in order to make them more clear, more easily understandable and to compare distributions quickly. It is also easy to understand by illiterate persons and people from different regions with different languages. Graphical representation brings to light the salient features of the data at a glance. It is also useful in locating some partition values.

The following graphs are generally used in representing frequency distributions :

1. Dot Frequency Diagram
2. Histogram and Bar Diagram

3. Frequency Polygon and Frequency Curve
4. Cumulative Frequency Curve or Ogive
5. Pie Chart

Dot Frequency Diagram:

The X-axis is used for the variable values and the Y-axis is for the frequency; if we indicate the frequencies of each variable value by dots, the resulting diagram is known as dot frequency diagram.

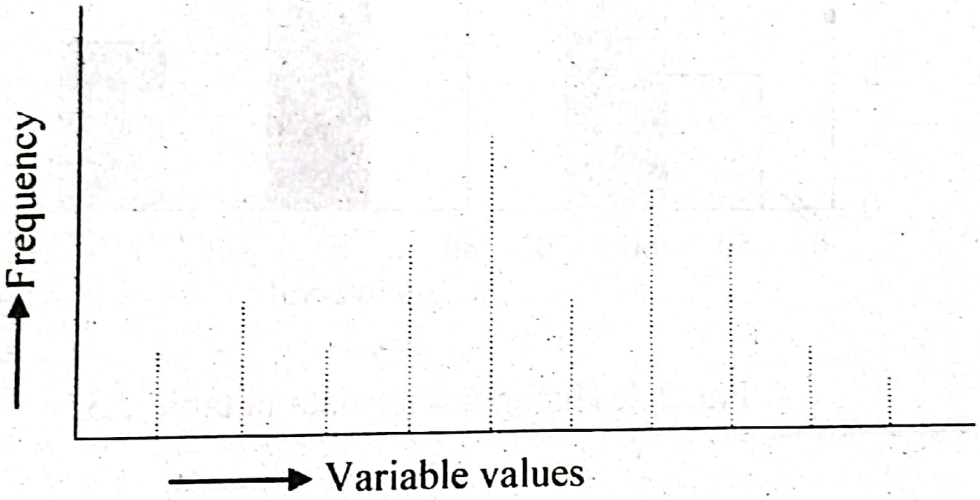


Fig. 2.1: Dot frequency diagram.

Histogram :

Class intervals are plotted along the X-axis and frequencies are plotted along the Y-axis. For each class or group, a rectangle is drawn taking class interval as the base and the class frequency as the height. For continuous variable, the rectangles such drawn are attached to adjacent rectangles at both sides and the resulting graph is known as histogram.

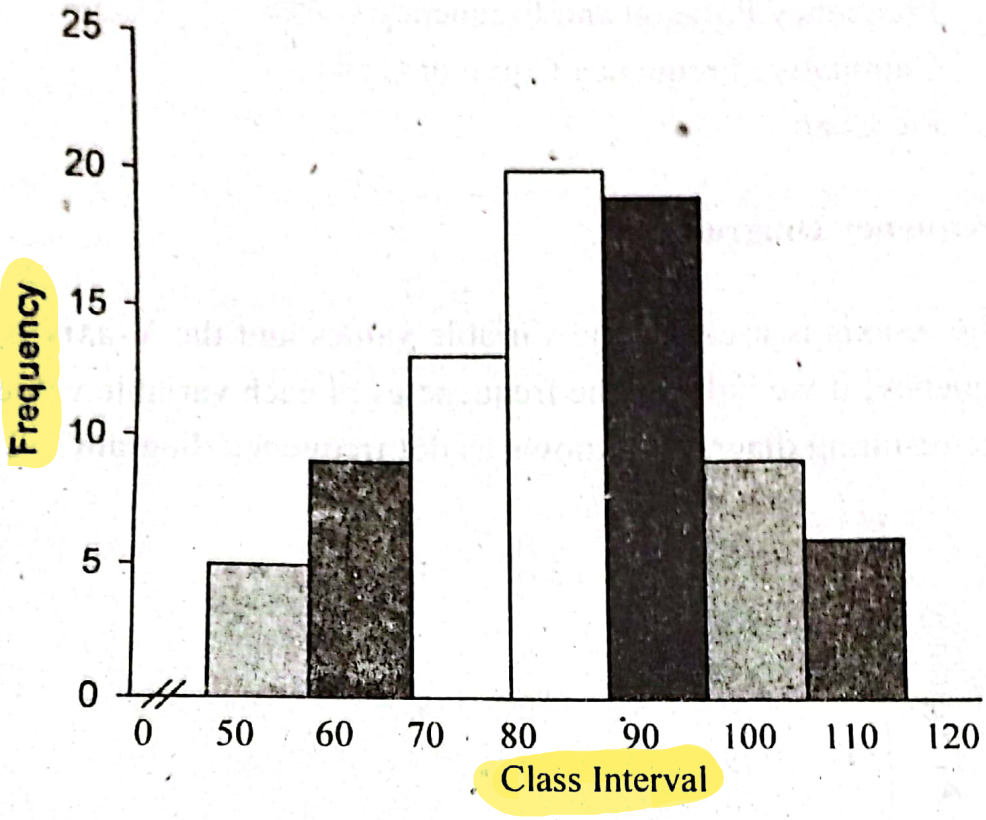


Fig. 2.2: Histogram for data in table 2.3.

For drawing histograms of frequency distributions havi unequal class intervals, frequency density, instead of frequency plotted along the Y-axis.

Frequency density is obtained as $f_d = \frac{f}{c}$; c being the cl interval.

Exam ple 2.4:

Drawing the histogram of the distribution of members per fam in a certain locality is described below :

Family size (Class Interval)	No. of families (f)	Class interval (C)	Frequency density $f_d = \frac{f}{c}$;
0-2	8	2	4
2-4	14	2	7
4-8	16	4	4
8-12	20	4	5
12-20	8	8	1
Total	66		

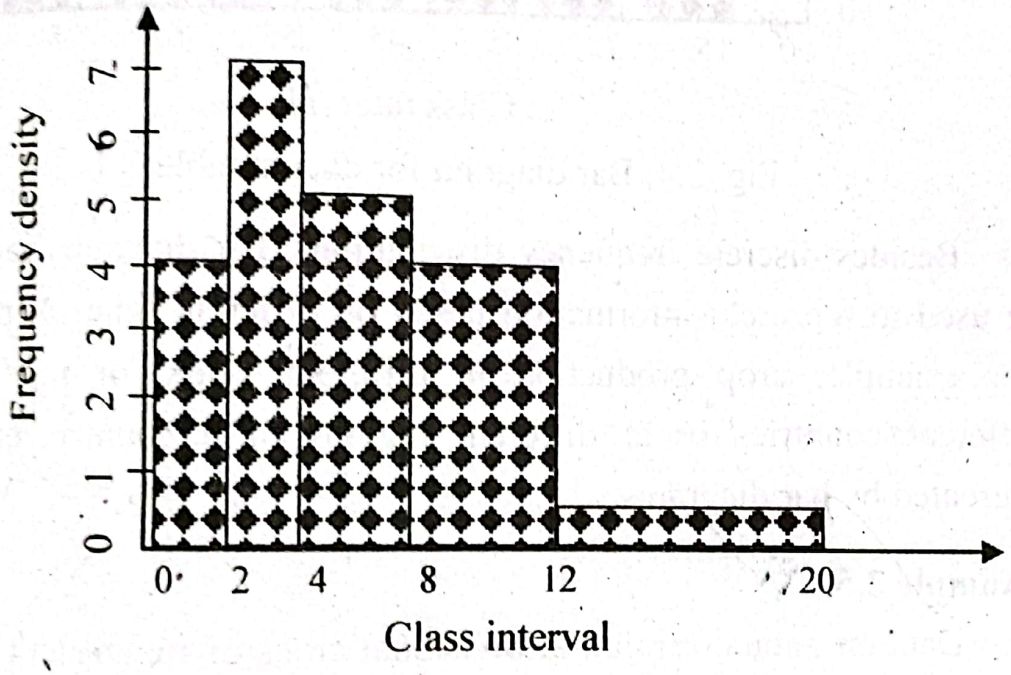


Fig. 2.3: Histogram for data with unequal class interval (Example 2.4)

Bar Diagram : Bar diagram is used mainly to represent discrete frequency distributions. Drawing process of bar diagram is similar to that of histogram. For discrete variables a gap exists between the upper limit of a class and the lower limit of the following class and the adjacent rectangles are not attached to each other. The graph is known as bar diagram.

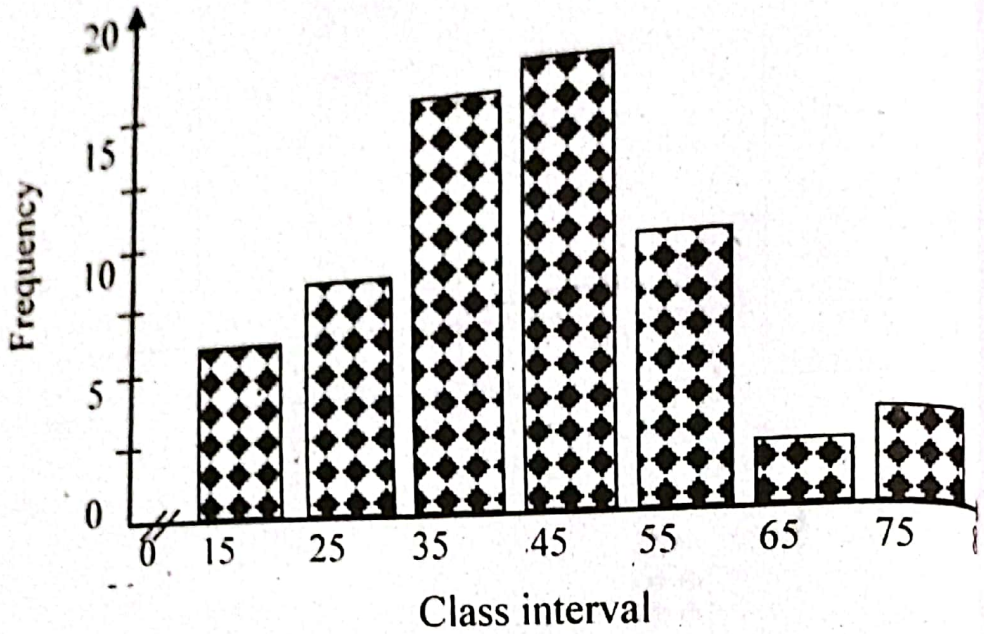


Fig. 2.4: Bar diagram for data in table 2.4.

Besides discrete frequency distributions, bar diagrams can be used to represent information based on different times or pl For example, crop production in different years, or rainfal different countries or at different regions of a country may presented by bar diagrams.

Example 2.5:

Data on annual rainfall at divisional cities of Bangladesh a follows :

Division	Annual Rainfall (mm.)
Dhaka	1540
Chittagong	2260
Khulna	1159
Rajshahi	1142
Sylhet	3568
Barisal	1200

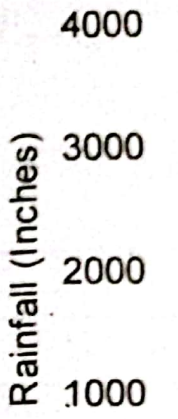


Fig.

Multiple Bar

Data on points may be for the varia constructed si the values of same place or identification.

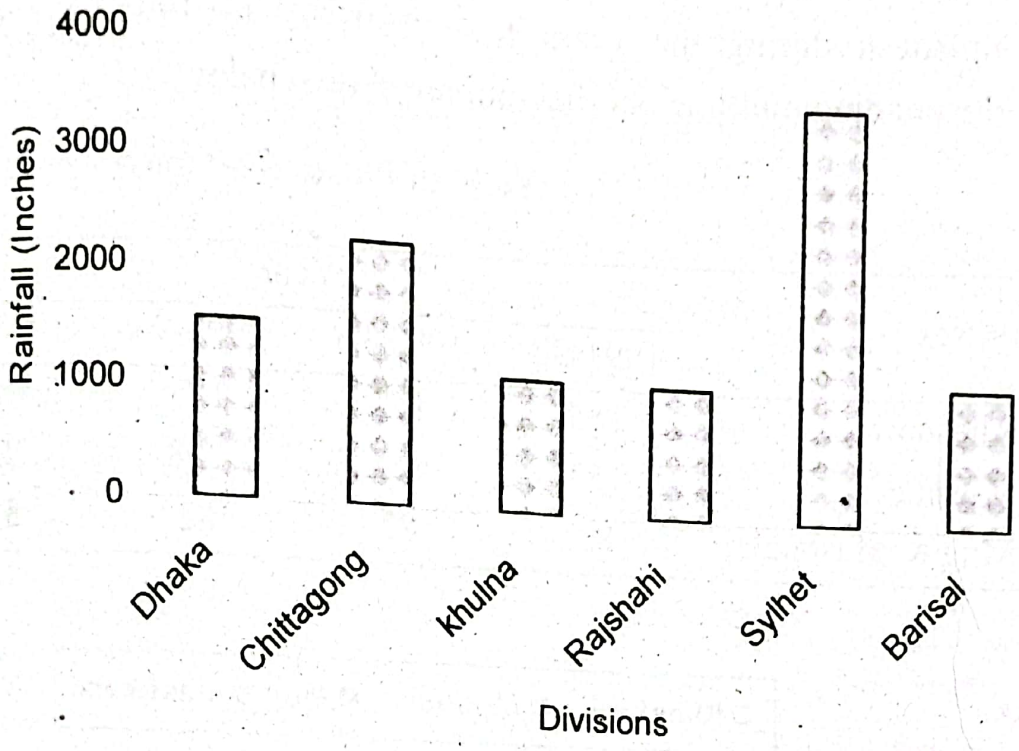


Fig. 2.5: Bar diagram of rainfall at different divisions

Multiple Bar Diagram :

Data on several variables in respect of different places or time-points may be represented by multiple bar diagrams. The simple bars for the variables corresponding to a place or time-point are constructed side by side (without gap). Heights of these bars indicate the values of the respective variables. Usually different bars at the same place or time-point are given with different colours or marks for identification.

Example 2.6:

Data on production of different pulses (in '000 tons) in Bangladesh during the years from 1991-92 to 1994-95 and the corresponding multiple bar diagram are shown below :

Yield of Pulses (000 tons)

Pulses	Year			
	1991-92	1992-93	1993-94	1994-95
Kheshari	185	172	188	189
Moshur	153	163	168	168
Mug & Mashkalai	82	82	82	85

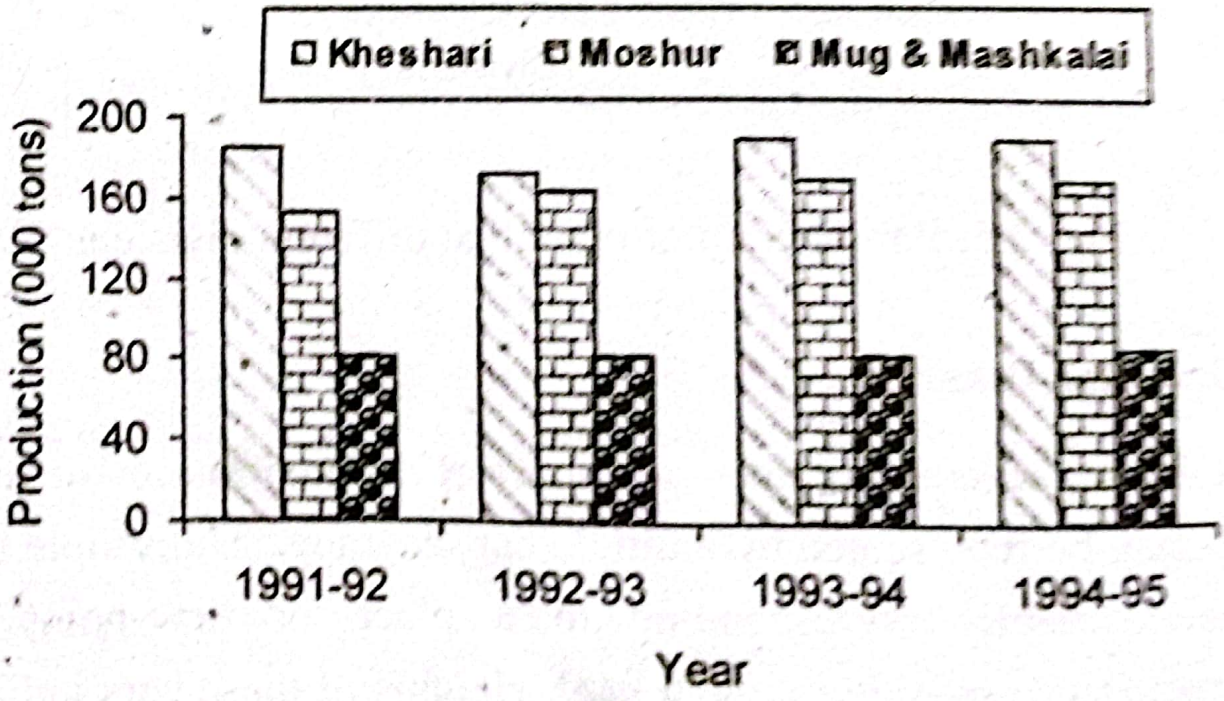


Fig. 2.6: Multiple bar diagram of pulses production.

Comparison between Histogram and Bar diagram

Histogram	Bar diagram
Histograms are used to represent continuous frequency distributions.	Bar diagrams are used to represent discrete frequency distributions.
Histograms are used to represent frequency distribution only.	Besides frequency distributions, data on different places or time points can be represented by bar diagrams.
In drawing the rectangles of the histogram both breath and length of the rectangles are considered.	In drawing bar diagrams consideration of the breath of bars is not necessary. For descent pictorial presentation, bars of suitable breath may be drawn.
Frequency distributions having unequal class intervals may be represented by histograms; in such case the breath of rectangles are unequal.	Bar diagrams are not usually drawn to represent frequency distributions having unequal class intervals.

Frequency Polygon and Frequency Curve :

Mid values of the class intervals are plotted along the X-axis and points are marked on the basis of respective class frequencies plotted along the Y-axis. If the consecutive points are connected with straight line, the resulting graph is a frequency polygon. At the starting and at the end of the distribution one class each are assumed and the polygon, for its completion, is extended upto the mid points of these class intervals, taking their frequencies to be zero.

Frequency polygon for table 2.3 is shown below :

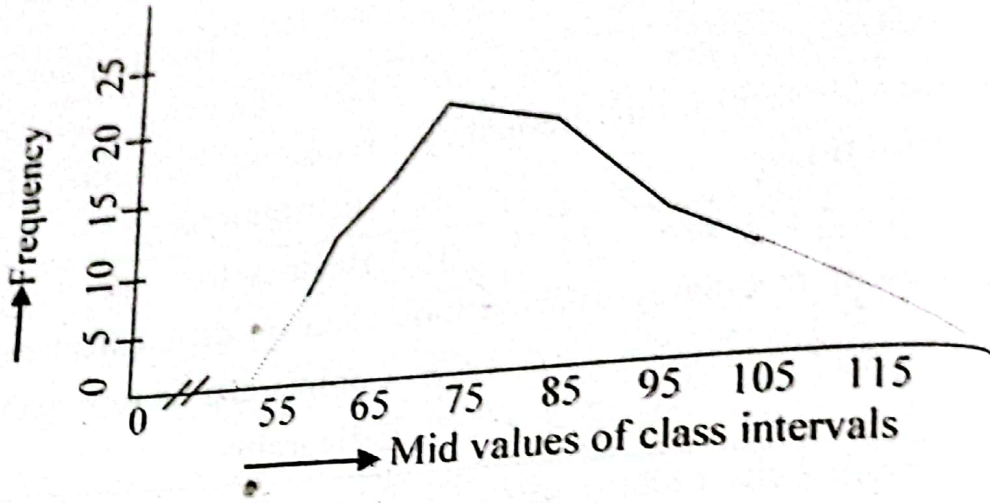


Fig. 2.7: Frequency polygon

Frequency polygon can also be obtained by joining the mid points of the vertical lines of the rectangles of the histogram.

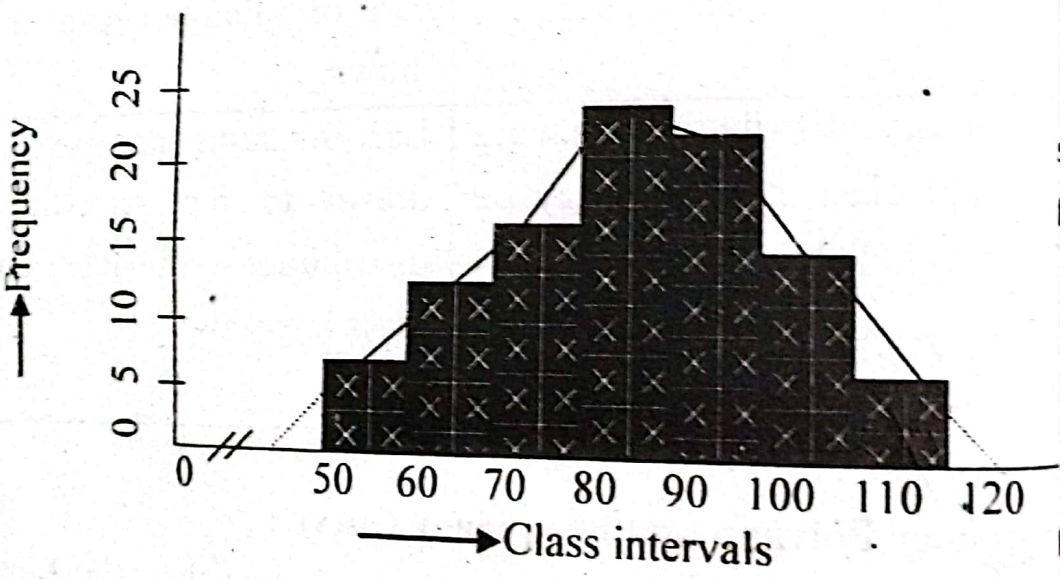


Fig. 2.8: Drawing frequency polygon from histogram in fig. 2.2

Frequency Curve :

In drawing frequency polygons, the consecutive points are connected by straight lines. If the points are connected by a free hand smooth curve, the resulting graph is known as frequency curve. A the frequency curve is a free hand curve, usually it does not pas through all the points; of course, it is desirable to touch the majori

number of points. The frequency curve for the table 2.3 is shown below :

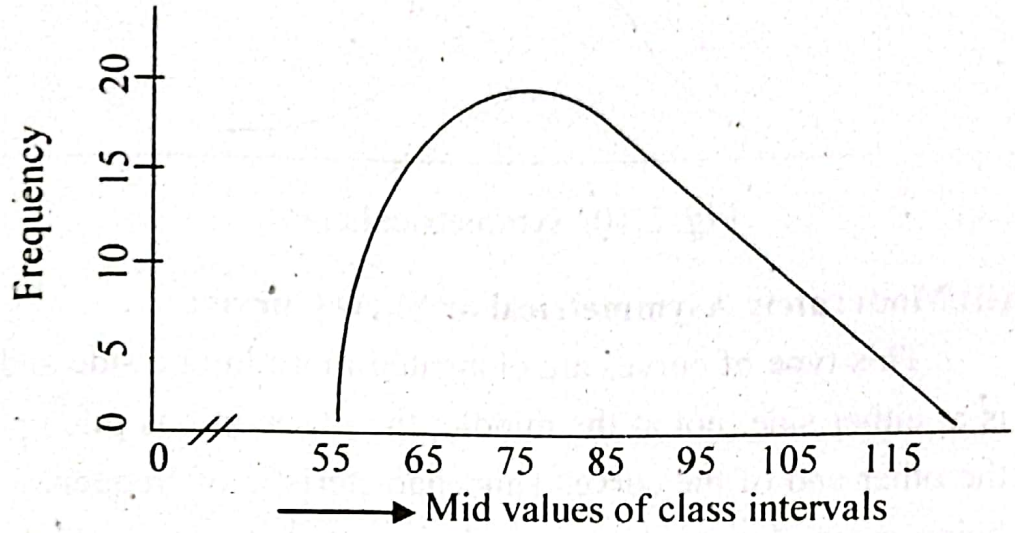


Fig. 2.9: Frequency curve

Forms of Frequency Curves :

The frequency curve gives an apparent idea of the nature and shape of the distribution. We generally get the following types of frequency curves in representing frequency distributions.

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

(i) Symmetrical Curve :

A symmetrical curve is one which has its peak at the middle point and gradually moves downwards in both sides with the same rate of decrease. The ordinates of the curve equidistant from its middle point are equal. If the curve is drawn on a paper and folded along a vertical line, the two-halves will coincide. An important type of a symmetrical curve which has got a single smooth hump in the middle and tails off gradually at both ends.

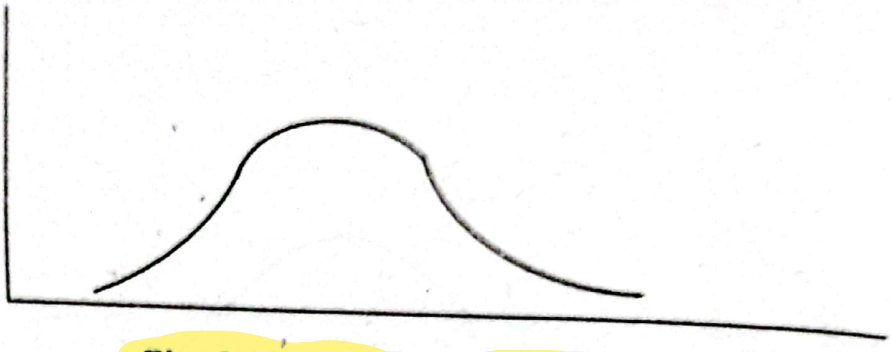


Fig. 2.10: Symmetrical curve

(ii) Moderately Asymmetrical or Skew Curve :

This type of curves are elongated more in one side and the p is at either side, not at the middle; the observations pile up at on the other end of the curve. This characteristic of frequency curve being more elongated at one side is called skewness. If the curv more elongated at the right side, it is said to be positively skewed.

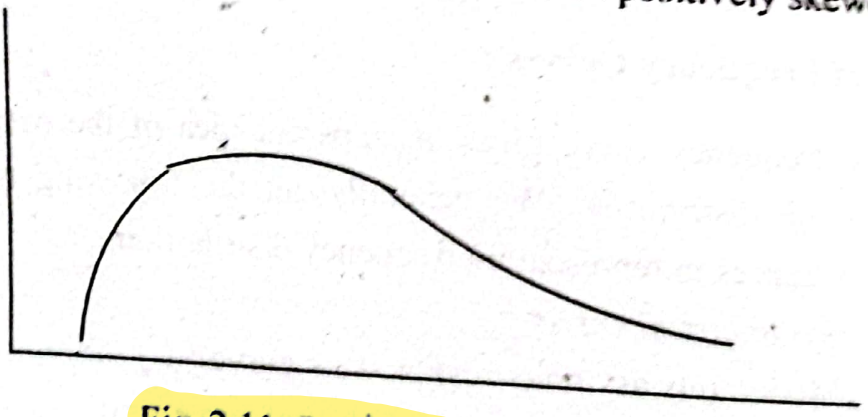


Fig. 2.11: Positively skewed curve

If the curve is more elongated at the left side, it is a negati skewed curve.

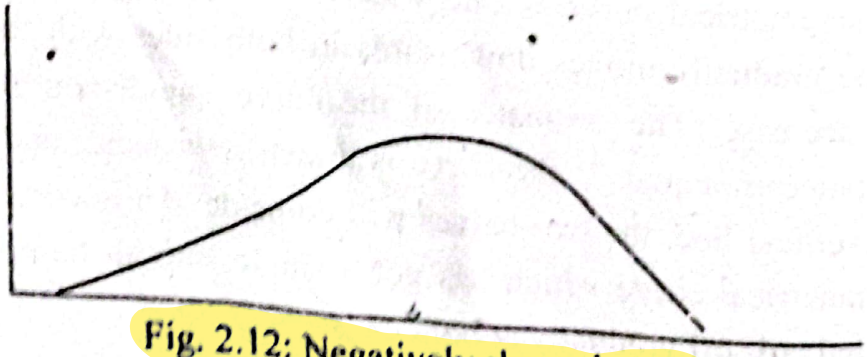


Fig. 2.12: Negatively skewed curve

(iii) Extremely Asymmetrical or J-Shaped Curve :

If the maximum frequency of a distribution occurs at one end and experiences a gradual decrease, the resulting frequency curve looks somewhat like the English letter J and hence it is often called J-shaped curve; the skewness is extreme. If the maximum frequency occurs at the starting, the curve is positively skewed and in the reverse case, it is negatively skewed. The curve of income distribution is an example of positive J-shaped curve.

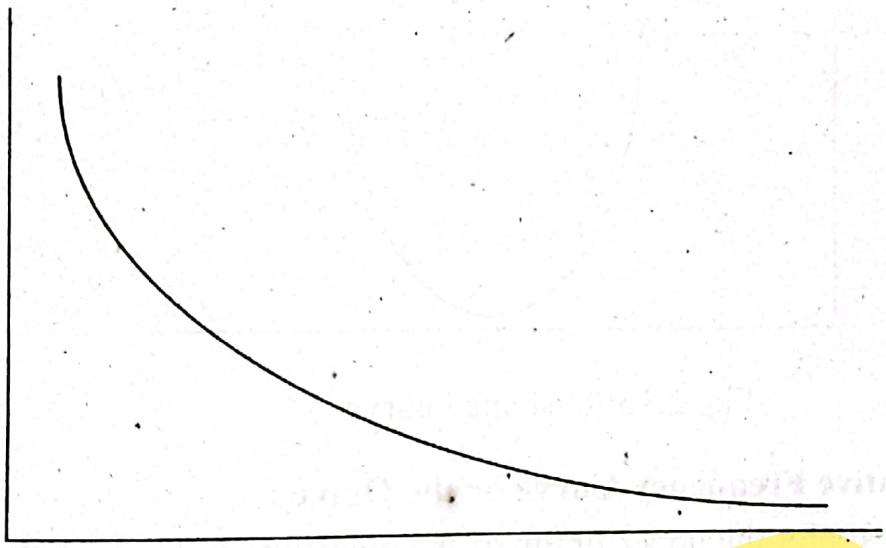


Fig. 2.13: Extremely asymmetrical (positive) curve

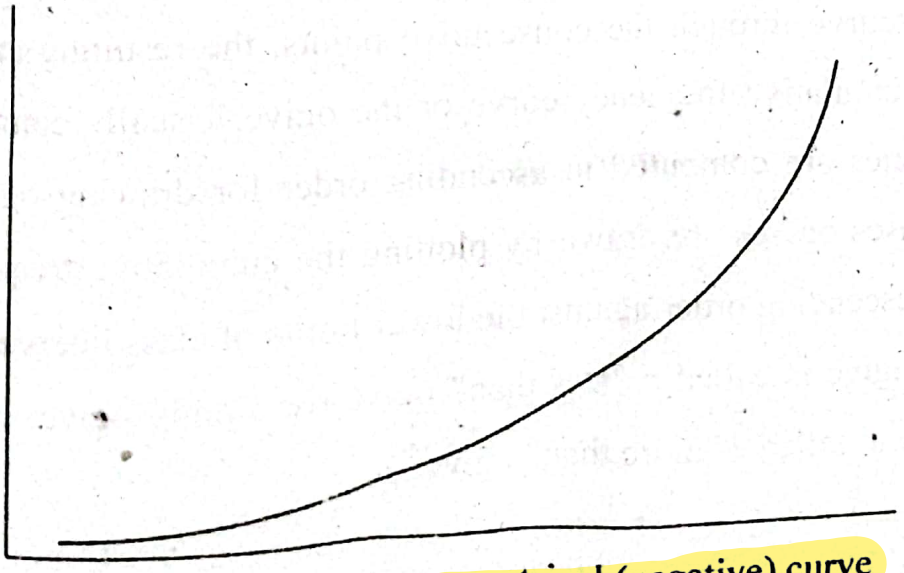


Fig. 2.14: Extremely asymmetrical (negative) curve

(iv) U-shaped Curve :

In some frequency distributions concentration of th observations are the maximum at both the ends and the minimum a the middle. The frequency curve of such a distribution looks like th letter "U" and hence it is called U-shaped curve. Frequenc distribution of age-specific human death rate exhibits a U-shape curve.

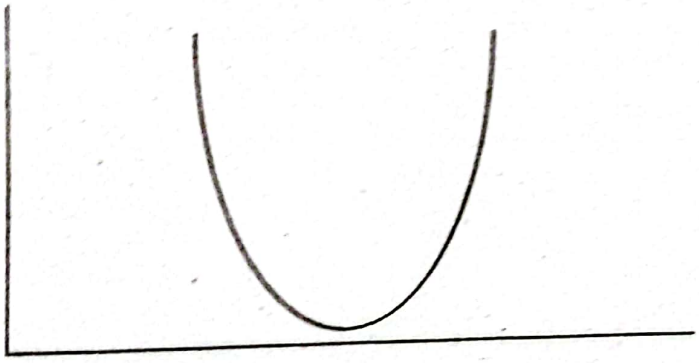


Fig. 2.15: U-shaped curve

Cumulative Frequency Curve or the Ogive :

If we plot the upper limits of the class intervals along the x-axi and the cumulative frequency along the y-axis and draw a free han smooth curve through the consecutive points, the resulting curve called cumulative frequency curve or the ogive. Usually cumulativ frequencies are computed in ascending order for drawing ogive. I some cases ogives are drawn by plotting the cumulative frequenci in the descending order against the lower limits of class intervals. Th former ogive is called a "less than" ogive (or simply ogive) and th later one is called a "more than" ogive.

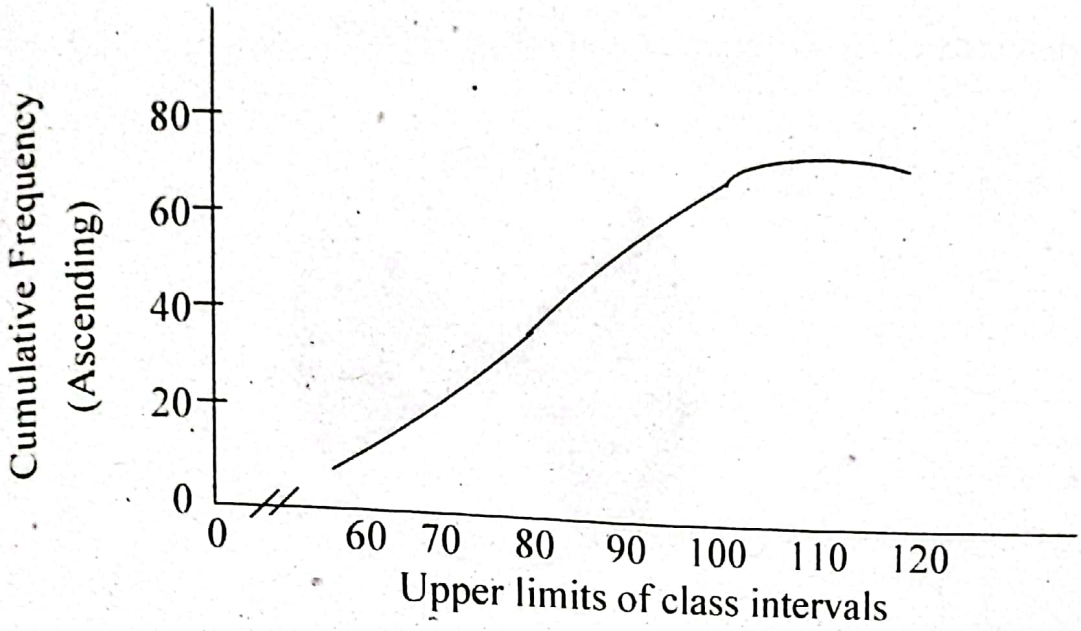


Fig. 2.16: Less than Ogive (or simply ogive) for the distribution in table 2.3

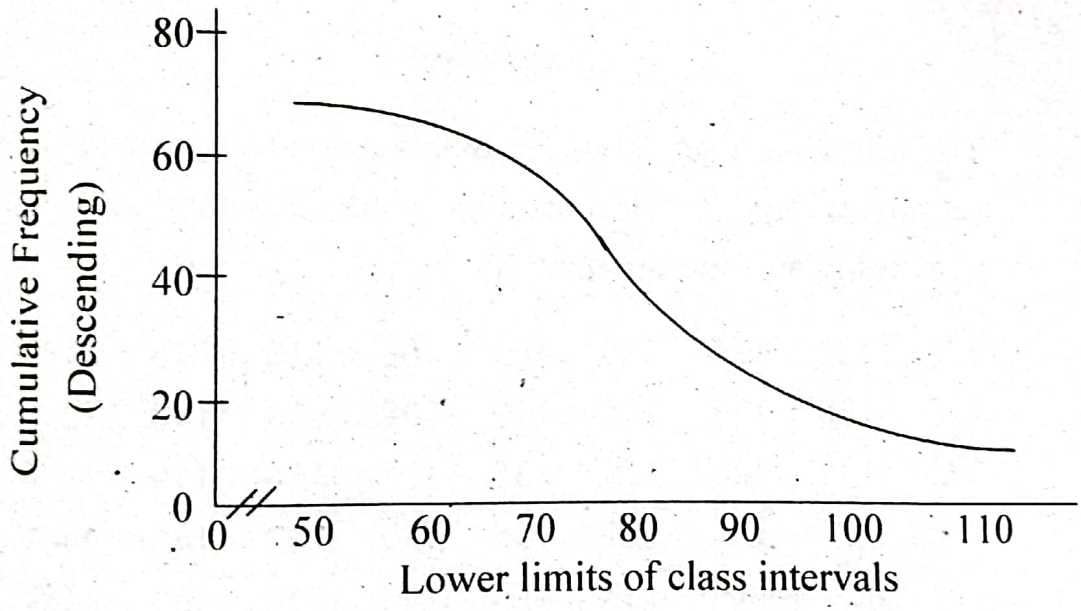


Fig. 2.17: More than ogive for table 2.3

Pie Chart :

Different components of data are exhibited by splitting a circle. The angle at the centre of a circle is proportionately divided and accordingly splitting the circle we exhibit different components of the data. The division is also done in percentage according to the relative

magnitude of different components. Usually different divisions are demarcated by different colours or symboles.

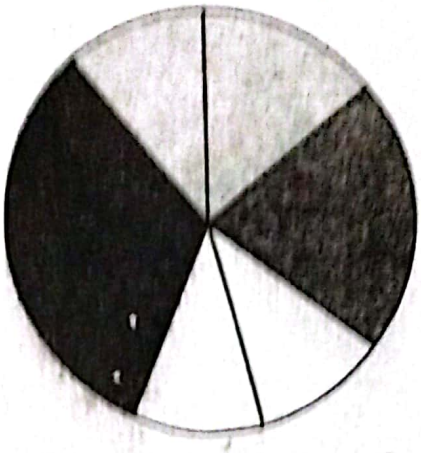


Fig. 2.18 : Pie chart of example 2.5

Exercises

1. The following data show the number of grains per panicle of a certain variety of rice; counts made on randomly selected panicles in an experimental plot :

135	120	125	140	145	105	110	115	118	128
125	130	128	115	110	108	138	140	129	126
102	135	126	129	135	128	126	119	122	130
126	129	132	121	136	139	142	145	148	107
132	127	119	128	130	145	126	103	118	128

a) Present the data in a frequency table taking a suitable class interval.

b) Draw a
(i) Bar diagram and
(ii) Frequency polygon

2. The weights (in gm) of apples selected at random from a consignment are given below :

125	111	123	95	82	109	93	115	107	187	76	107
105	92	70	126	186	100	115	123	119	139	129	78
110	78	84	99	141	111	104	136	123	90	115	82
90	107	81	131	75	84	165	152	188	170	155	148
92	85	93	105	90	99	110	108	80	82	130	122

a) Construct a frequency table taking a suitable class interval
b) Draw a (i) Histogram
(ii) Frequency polygon and
(iii) Less than ogive

3. The following frequency distribution shows the length of hilsha fish caught on a certain day at a certain point of the Padma.

Class Interval (Length in cm)	No. of Fishes caught	Class Interval (Length in cm)	No. of Fishes caught
25 - 30	39	45 - 50	15
30 - 35	45	50 - 55	8
35 - 40	52	55 - 60	5
40 - 45	75		

Draw (i) Histogram
(ii) Frequency curve and
(iii) Ogive

4. Marks obtained by 1st year Fisheries students in Statistics practical examination are given below: the full marks being 50.

Marks	No. of students	Marks	No. of students
20 - 22	5	28-30	15
22 - 24	10	30-32	4
24 - 26	12	32-34	4
26 - 28	28	34-36	2

Present the distribution by suitable graphs so as to make it more understandable.

5. Given below a frequency distribution of area (in sq.m.) of pond in a certain upazila of a certain district.

Area (sq.m.)	No. of ponds	Area (sq.m.)	No. of ponds
75 - 150	12	225 - 250	18
150 - 175	8	250 - 275	4
175 - 200	15	275 - 450	8
200 - 225	25		

Draw a histogram, a frequency polygon and an ogive.

6. The production of a certain crop in different years are given below:

Year	Production (000 'tons.)	Year	Production (000 'tons.)
1951	55	1981	72
1960	60	1991	58
1974	75	2001	40

Draw a bar diagram to represent the data.

7. Labour composition in selected types of activities in a country are given below :

Activity	Labour (%)		Activity	Labour (%)	
	Male	Female		Male	Female
Fertilizer factory	80	20	Food processing	30	70
Garments factory	15	85	Textile mills	65	35
Construction	70	30			
Sugar mills	75	25			

8. Composition of farmers according to farm size in a certain region of Bangladesh in different years are give below (Fictitious data) :

Year	Farmers (%)			Year	Farmers (%)		
	Small	Medium	Big		Small	Medium	Big
1975	30	35	35	1995	40	45	15
1980	32	40	27	2000	42	55	13
1985	35	30	35				
1990	38	42	20				

Draw multiple bar diagram.

9. Cost of living statistics of a group of service holders are given below :

Item	Monthly Expenditure (%)
Food	45
Cloth	20
Housing	25
Treatment	10

Draw a pie chart to present the expenditure pattern.

CHAPTER III

CENTRAL TENDENCY AND ITS MEASURES

The individual observations of a distribution or a data set are found to have a general tendency to cluster around a certain point, somewhere at the center of the distribution. For example, if we observe the distribution of the height of a group of students in a class, the height of most of the students are close to a certain central value. This tendency of the observations of a distribution to cluster or concentrate around the center of the distribution is called central tendency and its numerical measures are known as the measures of central tendency.

Different measures of central tendency are :

1. Mean
   - (a) Arithmetic mean (AM)
   - (b) Geometric mean (GM)
   - (c) Harmonic mean (HM)
2. Median and quantiles
3. Mode

The main purpose of measuring central tendency of a distribution is to determine such a value, which can be considered to be a representative one. An ideal measure of central tendency should, therefore, have the following characteristics :

- It should be rigidly defined
- It should be based on all the observations.
- It should be readily comprehensible and easy to calculate.
- It should be suitable for further algebraic treatment.
- It should be least affected by sampling fluctuations.

3.1 Arithmetic Mean :

Arithmetic mean of a set of observations is their sum divided by the number of observations.

The arithmetic mean may be of two types :

(a) Simple Arithmetic mean

(b) Weighted Arithmetic mean

(a) Simple Arithmetic Mean : The arithmetic mean $\bar{x}$ of n observations $x_1, x_2, \ldots, x_n$ is given by $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

In case of frequency distribution; let $f_1, f_2, \ldots, f_n$ are the frequencies of $x_1, x_2, \ldots, x_n$ respectively, arithmetic mean is obtained as

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

$$\left[\text{where : } \sum_{i=1}^{n} f_i = N\right]$$

(b) Weighted Arithmetic Mean : In practice all values of a series may not carry equal weight or importance. For example, if we want to have an idea of the change in cost of living of a certain group of people, the simple mean of prices of the commodities consumed by them will not do, since all the commodities are not equally important, e.g., rice, sugar and wheat are more important than confectionery items, coffee, tea, etc.

Let $w_1, w_2, \ldots, w_n$ be the weights attached to the item $x_1, x_2, \ldots, x_n$ respectively, the arithmetic mean is computed as-

$$\bar{x} = \frac{x_1w_1 + x_2w_2 \ldots + x_nw_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{W}$$

where $W = \sum_{i=1}^{n} w_i$

Properties of Arithmetic Mean :

1. **Sum of the deviations of the values of the variable from its** arithmetic mean is zero. i.e., $\sum_{i=1}^{n} f_i(x_i - \bar{x}) = 0$

Proof :

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

$$= N\bar{x} - \bar{x}\sum_{i=1}^{n} f_i = N\bar{x} - \bar{x}N = 0$$

Proved.

2. **Arithmetic mean is dependent on change of origin and scale.**

Proof : Let, $x_1, x_2, \ldots\ldots, x_n$ be the values of a variable x. Let us change the origin to an arbitrary value 'a' and change the scale by dividing by 'h'. The values of the new variable are, $u_i = \frac{x_i - a}{h}$

Now, $\bar{u} = \frac{1}{N}\sum_{i=1}^{n} f_i u_i = \frac{1}{N}\sum_{i=1}^{n} f_i\left(\frac{x_i - a}{h}\right)$

$$= \frac{1}{N}\frac{1}{h}\sum_{i=1}^{n} f_i(x_i - a) = \frac{1}{N}\frac{1}{h}\left\{\sum_{i=1}^{n} f_i x_i - \sum_{i=1}^{n} f_i a\right\}$$

$$= \frac{1}{N}\frac{1}{h}\{N\bar{x} - Na\} = \frac{1}{h}(\bar{x} - a)$$

or, $h\bar{u} = \bar{x} - a$

$\therefore \bar{x} = a + h\bar{u}$ Hence proved.

3. **The sum of the squares of the deviations of a set of values from their arithmetic mean is the minimum.**

Proof:

Let, $\bar{x}$ be the arithmetic mean of a set of observations $x_1, x_2, \ldots$ x_n with frequencies $f_1, f_2, \ldots\ldots, f_n$ respectively. Now sum of squares of deviations from an arbitrary value 'a' is

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

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

or, $\sum_{i=1}^{n} f_i (x_i - a)^2 = \Sigma f_i (x_i - \bar{x})^2 + n(x - a)^2$; $\quad [\because \Sigma f_i (x_i - \bar{x}) = 0]$

$\Rightarrow \sum_{i=1}^{n} f_i (x_i - a)^2 > \Sigma f_i (x_i - \bar{x})^2$; [Since $(x-a)^2$ is a positive quantity]

i.e., $\sum_{i=1}^{n} f_i (x_i - \bar{x})^2 < \sum_{i=1}^{n} f_i (x_i - a)^2$ Proved.

4. Mean of Composite Series :

If $\bar{x}_i$, (i = 1, 2, ... , k) are the means of k-component series of sizes n_i (i = 1, ..., k) respectively, then the mean $\bar{x}$ of the composite series can be obtained by the formula -

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

Proof:

Let $x_{11}, x_{12}, \ldots\ldots, x_{1n_1}$ be n_1 members of the first series,

$x_{21}, x_{22}, \ldots\ldots, x_{2n_2}$ be n_2 members of the 2nd series ;

.....

.....

$x_{k_1}, x_{k_2}, \ldots\ldots, x_{kn_k}$ be n_k members of the kth series-

having means $\bar{x}_1, \bar{x}_2, \ldots\ldots, \bar{x}_k$ respectively.

Then $n_1 + n_2 + \ldots\ldots + n_k$ will be the size of the composite series $(x_{11}, x_{12}, \ldots\ldots, x_{1n_1}), (x_{21}, x_{22}, \ldots\ldots, x_{2n_2}), \ldots\ldots, (x_{k_1}, x_{k_2}, \ldots\ldots, x_{kn_k})$.

The mean, $\bar{x}$ of the composite series of size $n_1 + n_2 + \ldots\ldots + n_k$ is given by

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

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

$$\therefore \bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i\bar{x}_i; \text{ where } N = \sum_{i=1}^{k} n_i \quad \text{Proved.}$$

Mean of first n Natural Numbers :

First n natural number are 1, 2, ..., n.

The mean, $\bar{x} = \frac{1 + 2 + \ldots\ldots + n}{n} = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$

$$\therefore \bar{x} = \frac{n+1}{2}$$

Advantages of Arithmetic mean:

- It is rigidly defined.
- It is easy to calculate.
- It is based upon all the observations.
- It is suitable for further algebraic treatment.
- It is less affected by sampling fluctuations.

Disadvantages of Arithmetic mean:

- It is affected very much by extreme values.
- It cannot be calculated if the extreme class is open.
- It is not suitable for extremely skewed distribution.
- It cannot be used if we are dealing with qualitative characteristics; such as intelligence, honesty, beauty, etc.
- It cannot be obtained if a single observation is missing or lost.

Uses of Arithmetic Mean:

- It is widely used to calculate average age, average income, average price, average salary, average increment, average import and average consumption, etc.
- It is used to establish the various theories and formulas of Mathematics and also used as an aid in further statistical analysis.
- It is used in computation of index number.

Example 3.1 : The daily wages of a group of farm workers are shown in the following frequency distribution.

Daily wages (Tk.)	Number of workers	Daily wages (Tk.)	Number of workers
50-55	5	70-75	15
55-60	10	75-80	7
60-65	25	80-85	3
65-70	35		

Computation of arithmetic mean by direct and indirect method.

Direct Method :

Daily wages (Tk.)	Number of workers f_i	Mid value x_i	$f_i x_i$
50-55	5	52.5	262.5
55-60	10	57.5	575.0
60-65	25	62.5	1562.5
65-70	35	67.5	2362.5
70-75	15	72.5	1087.5
75-80	7	77.5	542.5
80-85	3	82.5	247.5
	100		6640.0

$$\bar{x} = \frac{1}{N}\Sigma f_i x_i = \frac{1}{100}(6640.0) = \text{Tk. } 66.40$$

$\therefore$ Average daily wage is Tk. 66.40

Indirect Method :

Daily wages (Tk.)	Number of workers f_i	Mid value x_i	New variable $u_i = \frac{x_i - 67.5}{5}$	$f_i u_i$
50-55	5	52.5	-3	-15
55-60	10	57.5	-2	-20
60-65	25	62.5	-1	-25
65-70	35	67.5	0	0
70-75	15	72.5	1	15
75-80	7	77.5	2	14
80-85	3	82.5	3	9
	100			-22

New variable, $u_i = \frac{x_i - a}{h}$; where, a = 67.5 and h = 5

Now, $\bar{u} = \frac{1}{N}\Sigma f_i u_i = \frac{1}{100}(-22) = -0.22$

$\therefore \bar{x} = a + h\bar{u} = 67.5 + 5(-0.22) = 66.40$

$\therefore$ Average daily wage is Tk. 66.40

3.2. Geometric Mean (GM) :

Geometric mean of a set of n non-zero positive observations i the nth root of their product. The GM of n non-zero positive value $x_1, x_2, \ldots\ldots\ldots, x_n$ of a variable x is given by

$$GM = \sqrt[n]{x_1 . x_2 . \ldots\ldots x_n} = (x_1 . x_2 . \ldots\ldots\ldots . x_n)^{\frac{1}{n}}$$

$$\log (GM) = \log (x_1 . x_2 . \ldots\ldots x_n)^{\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$\therefore GM = \text{Anti log}\left\{\frac{1}{n}\sum_{i=1}^{n}\log x_i\right\}$$

In case of frequency distribution, when $f_1, f_2, \ldots\ldots, f_n$ be the frequencies of $x_1, x_2, \ldots\ldots, x_n$ respectively, then

$$GM = \sqrt[N]{x_1^{f_1}.x_2^{f_2}\ldots\ldots\ldots x_n^{f_n}}; \text{ where } N = \sum_{i=1}^{n} f_i$$

$$= \text{Anti log}\left[\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right]$$

(10 base)

Advantages of Geometric Mean :

- It is rigidly defined.
- It is based upon all the observations.
- It is not affected much by sampling fluctuations.
- It is suitable for further algebrical treatments.
- In measuring rate of change it is the most suitable average.

Disadvantages of Geometric Mean :

- It cannot be computed where there is any negative or zero values in the series.
- It is not easy to understand and to calculate for persons having very weak mathematical skills.
- It cannot be computed when the extreme classes of the frequency distribution are open.
- The value of GM may not be found in the series.

Uses of Geometric Mean :

- Geometric mean is used to find the average of ratios, rate of population growth, rate of interest, average of percentages.
- It is used in the construction of index numbers.

Example 3.2 :

Rate of increase of yield of a new wheat variety compared with a local variety in 10 selected agricultural farms are given below -

Rate of increase of yield (%)	Number of farm
0-5	1
5-10	2
10-15	4
15-20	2
20-25	1

For computation of geometric mean, we construct the following table

Rate of change of yield (%)	Frequency f_i	Mid value x_i	$\log x_i$	$f_i \log x_i$
0-5	1	2.5	0.39794	0.39794
5-10	2	7.5	0.87506	1.75012
10-15	4	12.5	1.09691	4.38764
15-20	2	17.5	1.24304	2.48608
20-25	1	22.5	1.35218	1.35218
	$\Sigma f_i = N = 10$			$\Sigma f_i \log x_i = 10.37396$

$$GM = \text{Anti log}\left\{\frac{1}{N}\Sigma f_i \log x_i\right\}$$

$$= \text{Anti log}\left\{\frac{1}{10}(10.373996)\right\} = \text{Anti log}\,(1.037396)$$

$$= 10.9 \text{ (Approx.)}$$

∴ The average rate of change of yield of the new variety of wheat is 10.9%

3.3 Harmonic Mean (HM) :

Harmonic mean of a set of non-zero observations is the reciprocal of the arithmetic mean of the reciprocals of the given values. Harmonic mean of n non-zero observations $x_1, x_2, \ldots\ldots, x_n$ is given by

$$HM = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}\right)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

$$\Rightarrow \frac{1}{HM} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

If $f_1, f_2, \ldots\ldots, f_n$ are respectively the frequencies of $x_1, x_2, \ldots\ldots, x_n$ non-zero observations; then

$$HM = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}(f_i / x_i)} = \frac{N}{\sum_{i=1}^{n}(f_i / x_i)}$$

$$\Rightarrow \frac{1}{HM} = \frac{1}{N}\sum_{i=1}^{n}(f_i / x_i)$$

Advantages of Harmonic Mean :

- It is rigidly defined.
- It is based upon all the observations.
- Sampling fluctuation is less.
- It is not affected much by extreme values

Disadvantages of Harmonic Mean :

- It cannot be computed where there is any zero values in the series.
- It is not easily understood and difficult to compute.
- It is very complex for further algebrical treatments.
- It cannot be computed if the extreme classes of the frequency distribution are open.

Uses of Harmonic Mean :

- The harmonic mean is used when observations are made in terms of work done per hour, speeds (kilometers covered per hour), quantity of things purchased per taka etc.

Example 3.3 :

The frequency distribution of profit per share of 10 companies are given below -

Profit per share (Tk).	0-5	5-10	10-15	15-20	20-25
No. of companies	1	2	4	2	1

To calculate harmonic mean, we construct the following table :

Profit per share (Tk.)	Frequency f_i	Mid value x_i	f_i / x_i
0-5	1	2.5	0.4000
5-10	2	7.5	0.2667
10-15	4	12.5	0.3200
15-20	2	17.5	0.1143
20-25	1	22.5	0.0444
Total	10		1.1454

$$HM = \frac{N}{\Sigma f_i / x_i} = \frac{10}{1.1454} = 8.73$$

$\therefore$ The average profit per share is Tk. 8.73

Relationship among AM, GM and HM:

- **For two non-zero positive observations:**

i) $A \geq G \geq H$

ii) $AH = G^2$; where A = Arithmetic mean,

G = Geometric mean and H = Harmonic mean.

Proof:

Let the two non-zero positive observations be x_1 and x_2.

By definition, $A = \frac{x_1 + x_1}{2}$; $G = \sqrt{x_1.x_2} = (x_1.x_2)^{1/2}$

and $H = \frac{1}{\frac{1}{2}\left(\frac{1}{x_1} + \frac{1}{x_2}\right)} = \frac{1}{\frac{x_1 + x_2}{2x_1x_2}} = \frac{2x_1x_2}{x_1 + x_2}$

i) Since any square quantity is always non-negative,

$$\therefore \left(\sqrt{x_1} - \sqrt{x_2}\right)^2 \geq 0$$

or, $x_1 + x_2 - 2\sqrt{x_1x_2} \geq 0$

or, $\frac{x_1 + x_2}{2} \geq \sqrt{x_1x_2}$ ..... (1)

$\therefore A \geq G$ ..... (2)

Again from equation (1)

$x_1 + x_2 \geq 2\sqrt{x_1x_2}$

or, $2\sqrt{x_1x_2} \leq x_1 + x_2$ or, $\frac{2\sqrt{x_1x_2}}{x_1 + x_2} \leq 1$

Multiplying both sides by $\sqrt{x_1x_2}$ we get

$$\frac{2\sqrt{x_1x_1}\sqrt{x_2x_2}}{x_1 + x_2} \leq \sqrt{x_1x_2}$$

or, $\frac{2x_1x_2}{x_1 + x_2} \leq \sqrt{x_1x_2}$

or, $H \leq G$ i.e., $G \geq H$ ..... (3)

From (2) & (3) it follows that $A \geq G \geq H$. Proved.

ii) $A.H = \frac{x_1 + x_2}{2} \times \frac{2x_1x_2}{x_1 + x_2} = x_1x_2$

$= \left(\sqrt{x_1x_2}\right)^2 = G^2$

$\therefore A.H = G^2$ Proved.

■ **For n non-zero positive observations:**

$$A \geq G \geq H$$

Proof. :

Let $x_1, x_2, x_3, \ldots\ldots, x_n$ be n non-zero positive observations.

Let, $d_i = x_i - A$

$\therefore x_i = A + d_i$ ..... (1)

By definition, $G = (x_1 . x_2 \ldots\ldots\ldots x_n)^{\frac{1}{n}}$

$$\Rightarrow \log G = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log(A + d_i) \; ; \qquad \text{From (1)}$$

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

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

$$= \log A + \frac{1}{n}\sum_{i=1}^{n} \{\frac{d_i}{A} - \frac{(d_i / A)^2}{2} \cdot \frac{1}{(1 + \theta d_i / A)^2}\};$$

[For $0 \leq \theta \leq 1$]

Expanding $\log(1 + d_i / A)$ in ascending power of (d_i / A) by Taylor's expansion method and assuming the higher order terms negligibly small, we get

$$\log G = \log A + \frac{1}{n}\frac{1}{A}\Sigma d_i - \frac{1}{n}\Sigma \frac{(d_i / A)^2}{2} \times \frac{1}{(1 + \theta d_i / A)^2}$$

Since, $\Sigma d_i = \Sigma(x_i - A) = 0$; as A = Arithmetic mean

and $(d_i / A)^2 . \frac{1}{(1 + \theta d_i / A)^2}$ is a positive quantity

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

$\Rightarrow \log A = \log G +$ a positive quantity

$\Rightarrow \log A \geq \log G$

$\therefore A \geq G$ ..... (2)

As $x_1, x_2, \ldots\ldots, x_n$ are non-zero positive, $\frac{1}{x_1}, \frac{1}{x_2}, \ldots\ldots, \frac{1}{x_n}$ are also non-zero positive and in this case also $A \geq G$

that is, $\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\ldots\ldots+\frac{1}{x_n}\right) \geq \sqrt[n]{\frac{1}{x_1} \times \frac{1}{x_2} \times \ldots\ldots \times \frac{1}{x_n}}$

or, $\frac{1}{H} \geq \frac{1}{\sqrt[n]{x_1 x_2 \ldots\ldots x_n}}$

or, $\frac{1}{H} \geq \frac{1}{G}$ $\quad \therefore G \geq H$ ..... (3)

From (2) and (3) it follows that

$A \geq G \geq H$ Proved.

3.4 Median :

The median of a distribution is the value of the variable which divides the distribution into two equal parts if arranged in order of magnitude. Median is the value such that the number of observations above it is equal to the number of observations below it. Thus median is the middlemost value of an ordered set and as such a positional average.

In case of ungrouped data, when the number of observations, n is odd, $\frac{n+1}{2}$th observation in the series will be the median. Again, when the number of observations, n is even, median will be the arithmetic mean of $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$th observation in the series.

For computing median from frequency distribution we first need to identify the median class (class which contains the median). If the total frequency is N, the class having cumulative frequency equal to, or immediately higher than N/2 will be the median class.

For frequency distribution, the formula for computing the median is

$$M_e = L_m + \frac{\frac{N}{2} - F_m'}{f_m} \times h$$

where,

L_m = lower limit of the median class.

N = total frequency

f_m = frequency of the median class

F_m' = cumulative frequency of the pre-median class

h = length of median class.

Advantages of Median :

- It is rigidly defined.
- It is easily understood and easy to compute.
- It is not influenced by extreme items.
- It can be calculated for distribution with opened classes.
- It can be used in defining the median of attributes.

Disadvantages of Median :

- It is not based upon all the observations.
- It is not suitable for further algebraic treatment.
- It is affected much by the sampling fluctuation.

Uses of Median :

- It is used in case of both quantitative and qualitative data.
- It is used for calculating the typical value in problem concerning wages, distribution of wealth etc.

3.5 Quantiles : Quantiles also are some positional or location measures of the distribution. Quantiles are those values in a series, which divide the whole distribution into a number of equal parts when the series is arranged in order of magnitude of observations. The following are the quantiles that are used in Statistics -

1) Quartiles (2) Deciles and (3) Percentiles.

3 quartiles : Q_i (i = 1, 2, 3); devide the whole distribution into four equal parts

9 Deciles : D_j (j = 1, 2,..., 9); devide the whole distribution into 10 equal parts.

99 Percentiles : P_k (k = 1, 2, ..., 99); devide the whole distribution into 100 equal parts.

Computation of quantiles from frequency distribution is very much similar to that of median. We first need to identify the corresponding quantile class. The classes having cumulative frequencies equal to or immediately higher than iN/4, jN/10 and kN/100 are respectively the ith quartile class, the jth decile class and the kth percentile class.

For frequency distributions the quantiles are computed as -

$$Q_i = L_i + \frac{\frac{iN}{4} - F_i'}{f_i} \times h ; \quad i = 1, 2, 3$$

$$D_j = L_j + \frac{\frac{jN}{10} - F_j'}{f_j} \times h ; \quad j = 1, 2, \ldots\ldots\ldots, 9$$

$$P_k = L_k + \frac{\frac{kN}{100} - F_k'}{f_k} \times h ; \quad k = 1, 2, \ldots\ldots\ldots, 99$$

i, j, k indicate the order of quartiles, deciles and percentiles respectively; F_i', F_j' and F_k' are respectively the cumulative frequencies of class preceding the ith quartile, jth decile and kth percentile classes; h is the corresponding class interval.

It may be mentioned that

$$Q_2 = D_5 = P_{50} = M_e ; \quad Q_1 = P_{25} ; Q_3 = P_{75} ; D_6 = P_{60} \text{ etc.}$$

Graphical Location of Median and Quantiles :

Median, quartiles, deciles and percentiles can be located from ogive; the necessary steps are briefly discussed below :

i) An ogive is drawn and the position in the Y-axis are marked for different partition values (e.g., $^{N}/_{2}$ for median, $^{N}/_{4}$ for 1st quartile, $^{4N}/_{10}$ for 4th decile etc.)

ii) From the corresponding points in the Y-axis, a line parallel to the X-axis is drawn which intersects the ogive at certain point.

iii) From the corresponding point of intersection mentioned above, a perpendicular is drawn on the X-axis; the foot of the perpendicular is the desired partition value. The whole process is illustrated in figure 3.1 below :

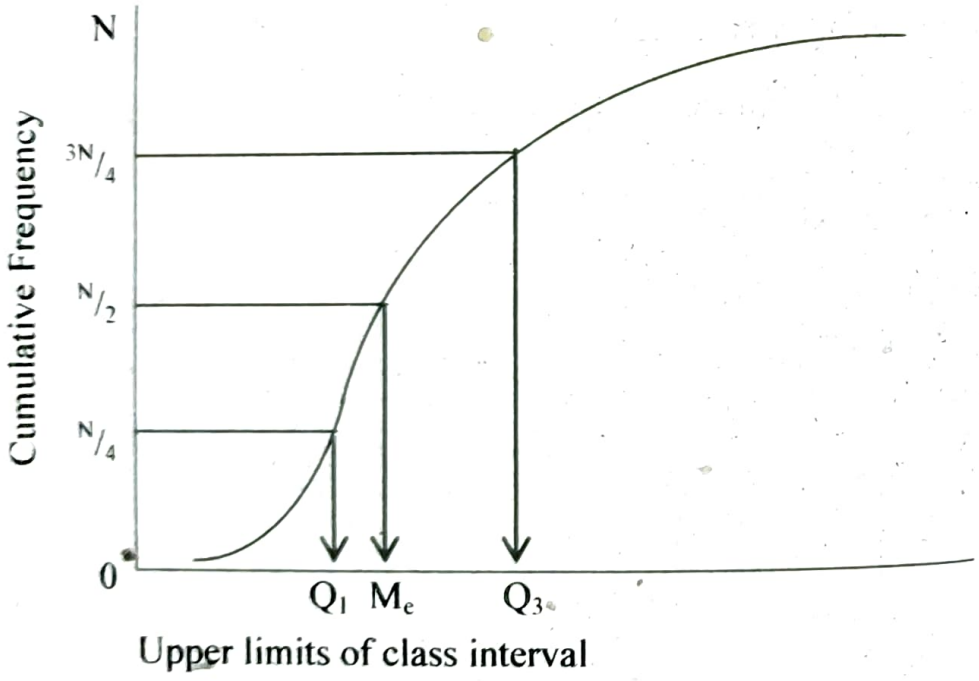


Fig. 3.1: Location of median and quantiles.

3.6. Mode (M_o) :

Mode of the distribution is that value of the variate for which the frequency is the maximum. In other words, mode is the highest

frequent value of a distribution. In the case of frequency distribution, mode is given by

$$M_o = L + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times h$$

where, L = lower limit of modal class
f_0 = frequency of modal class
f_1 = frequency of pre-modal class
f_2 = frequency of post-modal class

[The class which corresponds to the maximum frequency is the model class]

Advantages of Mode :

- It is easy to understand and easy to calculate
- It is not affected by extreme values.
- It can be located graphically.

Disadvantages of Mode :

- It is not rigidly defined - a distribution may have more than one mode.
- It is not based upon all the observations.
- It is not suitable for further algebraic treatment.

Uses of Mode :

- Mode is used to find the ideal size, e.g., in business forecasting, Meteorological forecast on weather condition, in the manufacture of ready-made garments, shoes, etc.

Graphical Location of Mode :

Mode can graphically located in two ways :

a) Using frequency curve.

b) Using the histogram.

a) From the peak of the frequency curve, a perpendicular is drawn on the X-axis; the foot of the perpendicular indicates the mode (shown in figure 3.2) :

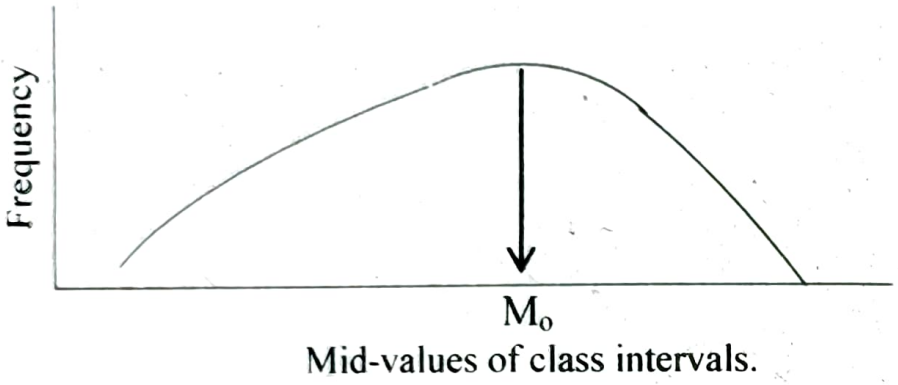


Fig. 3.2: Location of mode from frequency curve

b) Mode can be located more accurately from the histogram; the steps are the following :

i) The rectangles corresponding to the modal group, the pre-modal group and the post-modal group are considered. A straight line is drawn connecting the left vertical point (say A) of the modal group rectangle and the left vertical point (say D) of the post modal group rectangle. Similarly the right vertical point (say B) of the modal group rectangle and the right vertical point (say C) of the pre-modal group rectangle are connected.

ii) From the point of intersection of AD and BC, a perpendicular is drawn on the X-axis; the foot of the perpendicular indicates mode.

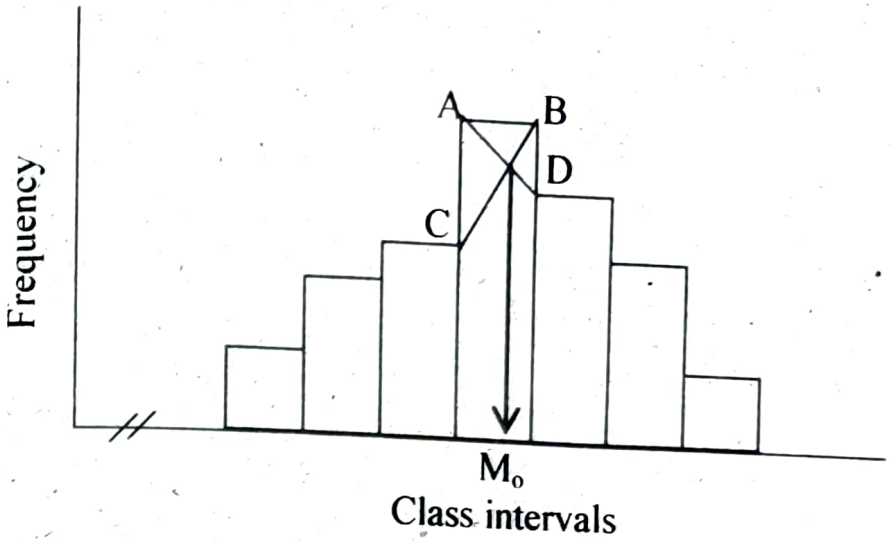


Fig. 3.3: Location of mode from the histogram.

Comparison Among the Measures of Central Tendency

Criteria	AM	GM	HM	M_e	M_o
Definition	Rigidly defined	Rigidly defined	Rigidly defined	Rigidly defined	Not rigidly defined
Data restriction	No restriction	Values must be nonzero and positive	Values must be nonzero	No restriction	No restriction
Compu-tation	Easy	Slightly difficult	Slightly difficult	Easy	Easy
Based upon all observa-tions	Yes	Yes	Yes	No	No
Effect of extreme values	Less affected	Not affected	Less affected	Not affected	Not affected
Sampling fluctua-tion	Little	Little	Little	Much	Much
Graphical location	Not possible	Not possible	Not possible	Possible	Possible
Further algebrical treatment	Possible	Possible	Not possible	Not possible	Not possible

From the above comparison, it is clear that arithmetic mean is the best measure of central tendency.

Example 3.4 :

Find the median, lower and upper quartiles, 4th decile, 70th percentile and mode for the following distribution :

Class:	50-60	60-70	70-80	80-90	90-100	100-110	110 and over
Frequency:	5	9	13	20	19	9	5

i) Draw the ogive and locate M_e, Q_3, D_4 and P_{70}

ii) Draw the Histogram and locate mode

Solution :

Class	Frequency	c.f.
50-60	5	5
60-70	9	14
70-80	13	27
80-90	20	47
90-100	19	66
100-110	9	75
110 and over	5	80

N = 80

Here, N = 80

Computation of Median :

$\frac{N}{2} = \frac{80}{2}$ = 40th observation lies in the class (80-90)

$\therefore$ (80-90) is the median class

$$\therefore M_e = L_m + \frac{N/2 - F_m'}{f_m} \times h$$

$$= 80 + \frac{40 - 27}{20} \times 10$$

$$= 86.5$$

$L_m = 80$
$N/2 = 40$
$F_m' = 27$
$f_m = 20$
$h = 10$

Computation of Quartiles :

$\frac{N}{4} = \frac{80}{4}$ = 20th observation lies in the class (70-80)

$\therefore$ (70-80) is the lower quartile (Q_1) class

$$\therefore Q_1 = L_1 + \frac{N/4 - F_1'}{f_1} \times h$$

$$= 70 + \frac{20-14}{13} \times 10$$

$$= 74.62 \text{ (app.)}$$

L_1	$= 70$
$N/4$	$= 20$
F_1'	$= 14$
f_1	$= 13$
h	$= 10$

Again, $\frac{3N}{4} = \frac{3(80)}{4}$ = 60th observation lies in the class (90-100)

$\therefore$ (90-100) is the upper quartile (Q_3) class

$$\therefore Q_3 = L_3 + \frac{\frac{3N}{4} - F_3'}{f_3} \times h$$

$$= 90 + \frac{60-47}{19} \times 10$$

$$= 96.84$$

L_3	$= 90$
$\frac{3N}{10}$	$= 60$
F_3'	$= 47$
f_3	$= 19$
h	$= 10$

Computation of Deciles :

$\frac{4N}{10} = \frac{4(80)}{4}$ = 32th observation lies in (80-90)

$\therefore$ (80-90) is the 4th deciles (D_4) class

$$\therefore D_4 = L_4 + \frac{\frac{4N}{10} - F_4'}{f_4} \times h$$

$$= 80 + \frac{32-27}{20} \times 10$$

$$= 82.50$$

L_4	$= 80$
$\frac{4N}{10}$	$= 32$
F_4'	$= 27$
f_4	$= 20$
h	$= 10$

Computation of Percentiles :

$\frac{70N}{100} = \frac{70(80)}{100}$ = 56th observation lies in the class (90-100)

$\therefore$ (90-100) is the 70th percentiles (P_{70}) class

$$\therefore P_{70} = L_{70} + \frac{\frac{70N}{100} - F'_{70}}{f_{70}} \times h$$

$$= 90 + \frac{56-47}{19} \times 10$$

$$= 94.74 \text{ (app.)}$$

L_{70}	$= 90$
$\frac{70N}{100}$	$= 56$
F'_{70}	$= 47$
f_{70}	$= 19$
h	$= 10$

Computation of Mode :

Here (80-90) is the modal class because maximum frequency (20) lies in that class

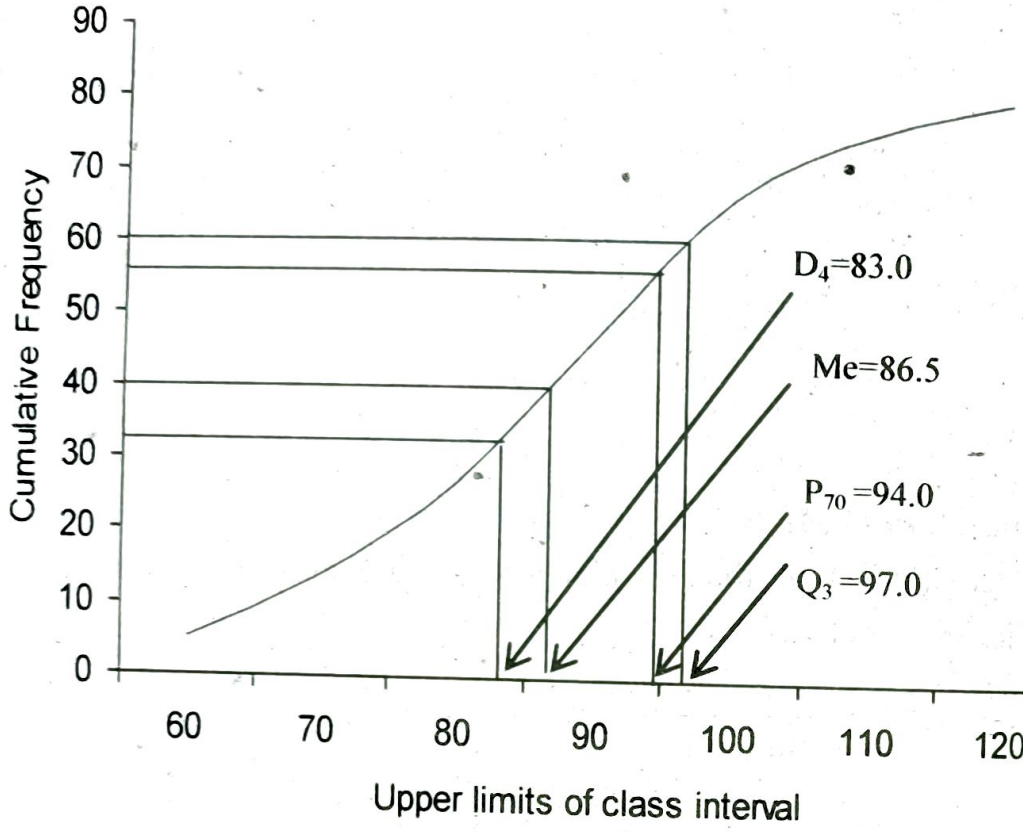
$$\therefore M_o = L + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times h$$

$$= 80 + \frac{20-13}{2 \times 20 - 13 - 19} \times 10$$

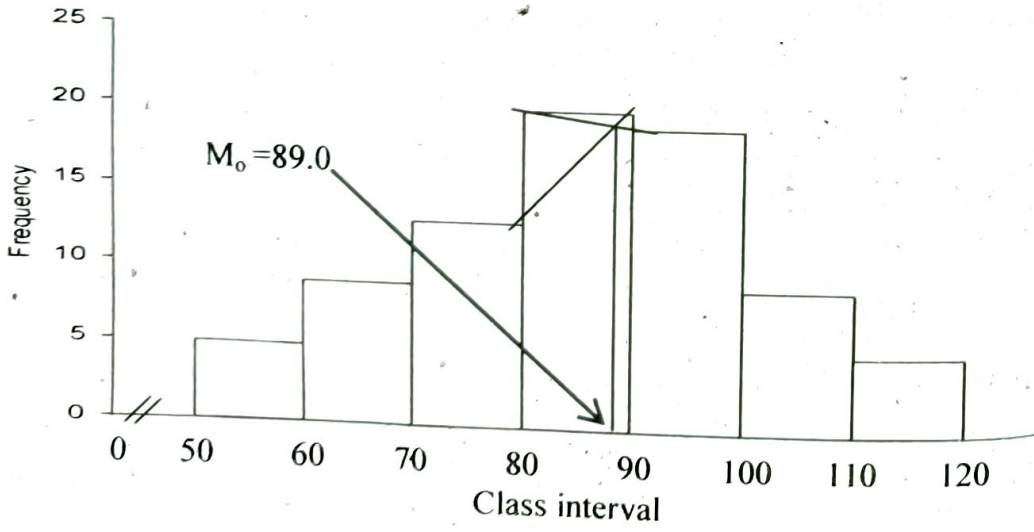
$$= 88.75$$

L	$= 80$
f_0	$= 20$
f_1	$= 13$
f_2	$= 19$
h	$= 10$

(i) Graphical location of quantiles from ogive :



ii) Location of mode from histogram.



Exercises

1. The rice yield (in kg) from a number of small plots are grouped with common class interval of 2 kg. in the table below; the x values are midvalues of the classes.

Yield(x)	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.2	4.4	4.6	4.8	5.0
No. of plots (f)	4	15	20	47	63	88	59	35	15	18	10	5

Compute arithmetic mean, geometric mean and harmonic mean and verify the relationship A.M. $\geq$ G. M. $\geq$ H.M.

2. The frequency distribution below gives the cost of production of sugar in different holdings :

Cost (Tk.)	No. of holdings)
10-14	11
14-18	27
18-22	42

Cost (Tk.)	No. of holdings)
22-26	45
26-30	35
30-34	30

Cost (Tk.)	No. of holdings)
34-38	20
38-42	15

Compute :

(a) A. M., G. M. and H. M.

(b) Median, Q_1, D_4, Q_3 and P_{80}.

(c) Mode

(d) Draw a histogram and locate the mode

(e) Draw a frequency curve and locate the mode.

(f) Draw an ogive and locate median, Q_1, D_4, Q_3 and P_{80}.

3. A set of 20 observations gives arithmetic mean 45 units and another set of 30 observations has arithmetic mean 60. The two set are combined; find the arithmetic mean of the combined set.

CHAPTER IV

DISPERSION, NATURE AND SHAPE OF FREQUEN(DISTRIBUTION

Central tendency is one character of a distribution. Measure central tendency give the idea of central value or location of distribution. But the central tendency is not the only character distribution. Two distributions may be different despite of their sa central value. As for example, the data set comprised of the value 10 and 20 has 10 as its mean and median. Again the mean and med of the series 5, 10, 15 is also 10. But the deviation of these val from their mean is not same. The deviation of observations from th mean is called dispersion. The measure of dispersion or variation the measure of the extent of variation or deviation of individ values from the central value. This measure of variation gives precise idea as to the extent of representativeness of the central valu

Characteristics of an Ideal Measure of Dispersion :

The following are the requisites for an ideal measure dispersion :

- It should be rigidly defined.
- It should be easy to understand and easy to calculate.
- It should be based on all the observations.
- It should be suitable for further algebrical treatments.
- It should be least affected by sampling fluctuation.
- It should be least affected by extreme values.

Importance of Measuring Dispersion :

Dispersion is an important character of distribution. Measures dispersion are widely used for the accurate and efficient analysis data. The importance of measures of dispersion are :

- Measure of dispersion is needed to know representativeness of the observations of a distribution; representativeness of mean can not be judged without the knowledge about dispersion.
- Measures of dispersion help to control the deviation of data.
- Measures of dispersion give the comparative picture of different distributions.
- Measures of dispersion help to control the quality of industrial products.
- Measures of dispersion is important for time series data such as rainfall, temperature etc., where central values are less important.

Measures of Dispersion may be divided in two broad types :

(a) Absolute Measures and
(b) Relative Measures.

(a) Absolute Measures :

1. Range 2. Quartile Deviation 3. Mean Deviation and 4. Standard Deviation

- Absolute measures of dispersion will retain the unit of measurement of the variable.

(b) Relative Measures :

1. Co-efficient of Range 2. Co-efficient of quartile deviation
3. Co-efficient of mean deviation 4. Co-efficient of variation.

- Relative measures of dispersion have no unit because these are the ratio of absolute measures and the corresponding values.

4.1. Absolute Measures of Dispersion :

Range :

Range is the absolute difference between the highest and low observations of a distribution. When the frequency distribution arranged in order of magnitude then range will be the absc difference between the mid-values of last class and first class.

Symbolically : Range = $| X_{max} - X_{min} | = | X_M - X_L |$

Range is the samplest and a crude measure of dispersion. Ra is based on two extreme observations only.

Advantages of Range :

- It is very easy to understand and easy to calculate.
- It gives us a quick idea about the variability of a set of data
- It is based on the extreme observations only and no det information is required.
- It is the simplest of all measures of distribution.

Disadvantages of Range :

- It is very much affected by the extreme values.
- It provides us with the idea of only two extreme values in set of data.
- It cannot be computed for data set having open ended cl interval.

Uses of Range :

- Range is used to forecast the weather, the percentage humidity in the air for weather forecasting.
- It is used in reporting daily market price of commodities.
- It is used in statistical quality control.

Quartile Deviation :

The quartile deviation is the half of the difference between upper quartile (Q_3) and lower quartile (Q_1).

$$QD = \frac{Q_3 - Q_1}{2}$$

It is also known as semi-interquartile range.

Advantages of Quartile Deviation :

- It is a very easily understandable location based measure.
- It is superior to other measures in the sense that the extreme values cannot affect the quartile deviation.
- For distributions with open ended class intervals no other measure can be computed but it is possible to compute quartile deviation.

Disadvantages of Quartile Deviation :

- It is not a good measure of dispersion because it does not measure the deviation from any central value of the distribution.
- It is not based upon all the observations.
- It is more affected by sampling fluctuations.
- It is not suitable for further algebric treatment.

Uses of Quartile Deviation :

- Quartile deviation is a location-based measure and can be profitably used where a rough estimate of the variation is desired.
- It is a suitable measure of dispersion when the frequency distribution has open-ended class interval.

Mean Deviation :

The arithmetic mean of the absolute deviations of the given observations from their central value is called mean deviation; it can be measured from mean, median and mode.

Mean deviation of a distribution having observations $x_1, x_2, \ldots\ldots x_n$ may be defined as follows :

- Mean deviation from mean or simply mean deviation :

$$MD(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

In the case of frequency distribution

$$MD(\bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - \bar{x}|; \text{ where } N = \sum_{i=1}^{n} f_i$$

- **Mean deviation from median :**

$$MD(M_e) = \frac{1}{n}\Sigma|x_i - M_e|$$

In the case of frequency distribution

$$MD(M_e) = \frac{1}{N}\Sigma f_i|x_i - M_e|$$

- **Mean deviation from mode :**

$$MD(M_o) = \frac{1}{n}\Sigma|x_i - M_o|$$

In the case of frequency distribution

$$MD(M_o) = \frac{1}{N}\Sigma f_i|x_i - M_o|$$

Advantages of Mean Deviation :

- It is based on all the observations
- It is rigidly defined and easy to understand.
- It is not affected by the extreme values
- It is suitable for comparative discussion.

Disadvantages of Mean Deviation :

- It cannot be computed for open-ended class intervals
- It is not amenable to father algebraic treatment.
- It is seldom used in statistical decision making.

Example 4.1 :

Computing mean deviation of the daily wages of a group of farm labours (given in example 3.1): The mean median and mode are respectively, $\bar{x} = 66.40$, $M_e = 66.43$, $M_o = 66.57$.

Mid value (x_i)	f_i	$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - M_e \rvert$	$\lvert x_i - M_o \rvert$	$f_i\lvert x_i - \bar{x} \rvert$	$f_i\lvert x_i - M_e \rvert$	$f_i\lvert x_i - M_o \rvert$
52.5	5	13.9	13.93	14.17	69.5	69.65	70.85
57.5	10	8.9	8.93	9.17	89.0	89.30	91.70
62.5	25	3.9	3.93	4.17	97.5	98.25	104.25
67.5	35	1.1	1.07	0.83	38.5	37.45	29.05
72.5	15	6.1	6.07	5.83	91.5	91.05	87.45
77.5	7	11.1	11.07	10.83	77.7	77.49	75.81
82.5	3	16.1	16.07	15.83	48.3	48.21	47.49
Total	100				512.0	511.40	506.6

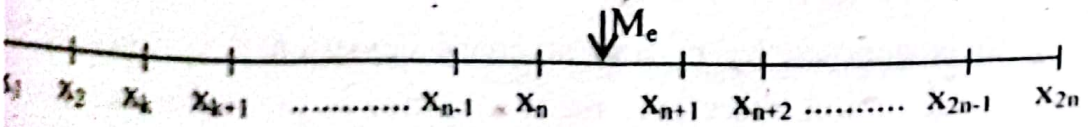
Mean deviation from mean $MD(\bar{x}) = \frac{1}{N}\sum f_i \lvert x_i - \bar{x} \rvert = \frac{512.0}{100} = 5.12$

Mean deviation from median $MD(M_e) = \frac{1}{N}\sum f_i \lvert x_i - M_e \rvert = \frac{511.4}{100} = 5.114$

Mean deviation from mode $MD(M_o) = \frac{1}{N}\sum f_i \lvert x_i - M_o \rvert = \frac{506.6}{100} = 5.066$

Theorem - 4.1: Mean Deviation from the Median is the Minimum.

Proof : Let 2n be the number of observations which are arranged in order of magnitude as $x_1, x_2, \ldots\ldots, x_k, x_{k+1}, \ldots\ldots, x_n, x_{n+1}, \ldots\ldots, x_{2n}$. Median ($M_e$) lies between x_n and x_{n+1} because observations are arranged in order of magnitude (shown below) :



Sum of the absolute deviations of observations from media given by

$$\Sigma |x_i - M_e| = [(M_e - x_1) + (M_e - x_2) + \ldots\ldots + (M_e - x_k)] + [(M_e - x_{k+1}) + (M_e - x_{k+2}) + \ldots\ldots + (M_e - x_n)] + [(x_{n+1} - M_e) + (x_{n+2} - M_e) + \ldots\ldots\ldots\ldots\ldots\ldots (1)$$
$$(x_{2n} - M_e)]$$

Again, sum of the absolute deviations of observations from a other value x_k is given by

$$\Sigma|x_i - x_k| = [(x_k - x_1) + (x_k - x_2) + \ldots + (x_k - x_k) + (x_{k+1} - x_k) + (x_{k+2} - x_k) + \ldots + (x_n - x_k)] + (x_{n+1} - x_k) + (x_{n+2} - x_k) + \ldots\ldots + (x_{2n} - x_k) \quad \ldots\ldots\ldots (2$$

Now, subtracting equation (1) from equation (2); we get

$$\Sigma |x_i - x_k| - \Sigma |x_i - M_e| = 2(n-k)\, x_{k+1} - 2(n-k)x_k$$
$$= 2(n - k)(x_{k+1} - x_k) \geq 0; [\text{Since } x_{k+1} >$$

$$\Rightarrow \Sigma |x_i - x_k| - \Sigma |x_i - M_e| \geq 0$$

$$\Rightarrow \Sigma |x_i - x_k| \geq \Sigma |x_i - M_e|$$

$$\text{or, } \frac{1}{2n}\Sigma|x_i - x_k| \geq \frac{1}{2n}\Sigma |x_i - M_e|$$

$$\text{or, } MD(x_k) \geq MD(M_e)$$

$$\text{i.e., } MD(M_e) \leq MD(x_k) \text{ ; for all } 0 < x < 2n$$

[Note : This theorem is always true for ungrouped data but may always be true for grouped frequency distribution; Ref: example above].

Standard Deviation :

The arithmetic mean of the squares of deviations of observations of a series from the their mean is known as variance. The positive square root of variance is called standard deviation. The variance is denoted by σ^2 and standard deviation is denoted by σ. Standard deviation, therefore, may be defined as the root mean square deviation from the mean.

For a set of observations $x_1, x_2, \ldots\ldots x_n$ standard deviation is computed as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

For frequency distributions.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2} \; ; \text{ where } N = \sum_{i=1}^{n} f_i$$

Root mean-square deviation from an arbitrary value a is denoted by s and is computed as $s = \sqrt{\frac{1}{N}\sum f_i(x_i - a)^2}$ $a = ?$

Standard Error : The standard deviation of the sampling distribution of a statistic (say mean) is known as standard error. It is denoted by SE.

Let $x_1, x_2, \ldots\ldots, x_n$ be the observations of a sample of size n. The standard error of mean is given by

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

σ = Population standard deviation, $\bar{x}$ = Sample mean (statistic)

Advantages of Standard Deviation :

- It is rigidly defined.
- It is based upon all the observations.
- It is less affected by sampling fluctuation.
- It is suitable for further algebric treatments.
- The standard deviation of the combined series can be obtained if the number of observations, mean and standard deviation in each series are known.

Disadvantages of Standard Deviation :

- It is not readily comprehensible, computation requires a good deal of time and knowledge of mathematics.

- It is affected by the extreme values.
- It cannot be computed in case of distributions having open ended class interval.

Uses of Standard Deviation :

Standard deviation is the most useful measure of dispersion. The use of standard deviation is highly desirable in advanced statistical works. Sampling and analysis of data have got their basis on standard deviation. Sampling, correlation analysis, the normal curve of errors, comparing variability and uniformity of two sets of data which are of great use in statistical works, are analysed in terms of standard deviation.

Thus standard deviation is the most important measure of dispersion.

Difference between Mean Deviation and Standard Deviation.

- In computing mean deviation (MD), we omit the sign of deviation but in computing standard deviation (s.d.) we do not need to omit the sign of the deviations.
- M.D. can be computed from mean, Median or Mode but in computing s.d. we consider only the deviations from the mean.
- M.D. is not suitable for further algebraic treatment but s.d. suitable for further algebraic treatment.

Some Properties of Standard Deviation :

1. Standard deviation is independent of change of origin but not of scale.
2. Standard deviation is the least possible root mean square deviation.
3. For two observations, standard deviation is the half of the range.

1. Standard Deviation is Independent of Change of Origin but not of Scale.

Proof.

Let, $x_1, x_2, \ldots\ldots, x_n$ be the mid-values of the classes of a frequency distribution and let $f_1, f_2, \ldots\ldots, f_n$ be their corresponding frequencies and also let, $u_i = \frac{x_i - a}{h}$; where u_i, a and h are changed variate, origin and scale respectively.

$$u_i = \frac{x_i - a}{h} \Rightarrow \bar{u} = \frac{\bar{x} - a}{h}$$

Now standard deviation of the new variable u is

$$\sigma_u = \sqrt{\frac{1}{N} f_i (u_i - \bar{u})^2}$$

$$= \sqrt{\frac{1}{N} \sum_{i=1}^{n} f_i \left\{ \frac{x_i - a}{h} - \frac{\bar{x} - a}{h} \right\}^2}$$

$$= \sqrt{\frac{1}{N} \sum_{i=1}^{n} f_i \left\{ \frac{x_i - a - \bar{x} + a}{h} \right\}^2}$$

$$= \sqrt{\frac{1}{N} \sum_{i=1}^{n} f_i \left(\frac{x_i - \bar{x}}{h} \right)^2} = \frac{1}{h} \sqrt{\frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x})^2}$$

$$= \frac{1}{h} \sigma_x$$

$$\Rightarrow \sigma_x = h \, \sigma_u$$

This implies that standard deviation is independent of change of origin but not of scale.

2. Standard Deviation is the Least Possible Root Mean Square Deviation.

Proof:

Let, $x_1, x_2, \ldots\ldots, x_n$ are the values of 'n' observations with corresponding frequencies $f_1, f_2, \ldots\ldots, f_n$. Also let $\bar{x}$ be the arithmetic mean of the observations.

We have, $\sigma_x = \sqrt{\frac{1}{N}\Sigma f_i (x_i - \bar{x})^2}$ and $s = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - a)^2}$

Mean square deviation from an arbitrary value 'a' is given by

$$s^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - a)^2$$

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

$$= \Sigma f_i (x_i - \bar{x})^2 + 2\Sigma f_i (x_i - \bar{x})(\bar{x} - a) + \Sigma f_i (\bar{x} - a)^2$$

$$= N\sigma_x^2 + 2(\bar{x} - a)\,\Sigma f_i (x_i - \bar{x}) + \Sigma f_i (\bar{x} - a)^2$$

$$= N\sigma_x^2 + 2(\bar{x} - a) \times 0 + \text{Positive quantity} \quad [\because \Sigma f_i (x_i - \bar{x}) = 0]$$

$\therefore Ns^2 = N\sigma_x^2 +$ positive quantity

$\therefore Ns^2 \geq N\sigma_x^2$

$\Rightarrow s^2 \geq \sigma_x^2$

i.e., $\sigma \leq s$

Proved.

3. For two Observations, Standard Deviation is the half of the Range.

Proof:

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

We have, $\sigma^2 = \frac{1}{2}\sum_{i=1}^{2}(x_i - \bar{x})^2 = \frac{1}{2}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2\}$

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

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

$$\Rightarrow, \sigma^2 = \left(\frac{x_1 - x_2}{2}\right)^2 = \left\{\frac{|x_1 - x_2|}{2}\right\}^2$$

$$\therefore \sigma = \frac{|x_1 - x_2|}{2} = \frac{1}{2}|x_1 - x_2| = \text{Half of range.}$$

- **Working Formula of Standard Deviation:**

Here,

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

$$= \sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2$$

$$= \sum x_i^2 - 2\left(\frac{\sum x_i}{n}\right)(\sum x_i) + n\left(\frac{\sum x_i}{n}\right)^2$$

$$= \sum x_i^2 - 2\frac{(\sum x_i)^2}{n} + \frac{(\sum x_i)^2}{n}$$

$$= \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$\therefore \sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$= \frac{1}{n}\left\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right\}$$

$$= \frac{1}{n}\sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2$$

$$\therefore \sigma = \sqrt{\frac{1}{n}\sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2}$$

In case of grouped data

$$\sigma = \sqrt{\frac{1}{N}\sum f_i x_i^2 - \left(\frac{\sum f_i x_i}{N}\right)^2}; \quad \text{where } N = \sum_{i=1}^{n} f_i$$

$$= \sqrt{\frac{1}{N}\left\{\sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{N}\right\}}$$

Theorem 4.2 : Standard Deviation is Smaller Than Range.

Proof :

Let, $\bar{x}$ = Mean and R = Range of 'n' observations $x_1, x_2, \ldots\ldots$

Since, Range is the difference between the highest and low observations of the distribution, it will be greater than $(x_i - \bar{x})$

i.e., $R > (x_i - \bar{x})$

We have, $\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$

$$= \frac{1}{n}\left\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots\ldots + (x_n - \bar{x})^2\right\}$$

$$< \frac{1}{n}\left\{R^2 + R^2 + \ldots\ldots + R^2\right\} = \frac{nR^2}{n} = R^2$$

$\therefore \sigma^2 < R^2$

$\therefore \sigma < R$ Proved.

Theorem 4.3 : Standard Deviation is Greater than Mean Deviation from Mean.

Proof :

By definition, $\sigma^2 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^2$ and $MD(\bar{x}) = \frac{1}{N}\Sigma f_i |x_i - \bar{x}|$

Let, $|x_i - \bar{x}| = z$

Now, $(z_i - \bar{z})^2 \geq 0$; Since any square quantity is positive.

$$\Rightarrow \frac{1}{N}\Sigma f_i(z_i-\bar{z})^2 \geq 0$$

$$\Rightarrow \frac{1}{N}\Sigma f_i(z_i^2+\bar{z}^2-2z_i\bar{z}) \geq 0$$

$$\Rightarrow \frac{1}{N}\Sigma f_i z_i^2+\frac{1}{N}\Sigma f_i\bar{z}^2-\frac{1}{N}2\Sigma f_i z_i\,\bar{z} \geq 0$$

$$\Rightarrow \frac{1}{N}\Sigma f_i z_i^2+\bar{z}^2-2\bar{z}^2 \geq 0$$

$$\Rightarrow \frac{1}{N}\Sigma f_i z_i^2-\bar{z}^2 \geq 0$$

$$\Rightarrow \frac{1}{N}\Sigma f_i z_i^2-\left(\frac{1}{N}\Sigma f_i z_i\right)^2 \geq 0$$

$$\Rightarrow \frac{1}{N}\Sigma f_i z_i^2 \geq \left(\frac{1}{N}\Sigma f_i z_i\right)^2$$

$$\Rightarrow \frac{1}{N}\Sigma f_i|x_i-\bar{x}|^2 \geq \left(\frac{1}{N}\Sigma f_i|x_i-\bar{x}|\right)^2$$

$$\Rightarrow \sqrt{\frac{1}{N}\Sigma f_i(x_i-\bar{x})^2} \geq \frac{1}{N}\Sigma f_i|x_i-\bar{x}| \qquad \left[\because |x_i-\bar{x}|^2=(x_i-\bar{x})^2\right]$$

$\therefore\ \sigma \geq MD(\bar{x})$ Proved.

- **Standard Deviation of first 'n' Natural Numbers.**

First n natural numbers are 1, 2, 3, ... , n.

$$\text{Variance, } \sigma^2 = \frac{1}{n}\left\{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right\}$$

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

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

$$= \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}\right]$$

$$= \left\{\frac{2n(n+1)(2n+1) - 3n(n+1)^2}{12n}\right\}$$

$$= (n+1)\left\{\frac{2n(2n+1) - 3n(n+1)}{12n}\right\}$$

$$= (n+1)\left(\frac{n^2 - n}{12n}\right) = (n+1)\frac{n(n-1)}{12n}$$

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

$$\therefore \sigma = \sqrt{\frac{n^2 - 1}{12}}$$

- **Standard Deviation of Combined Series.**

Let, x_{1i} $(i = 1, 2, \ldots\ldots, n_1)$ and x_{2j} $(j = 1, 2, \ldots\ldots, n_2)$ are two set with means $\bar{x}_1$ and $\bar{x}_2$ and variances σ_1^2 and σ_2^2 respectiv

Mean of the combined series is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{N}; \text{ where } N = n_1 + n_2$$

By definition,

$$\sigma_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 \text{ and } \sigma_2^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2$$

$$\therefore n_1\sigma_1^2 = \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_2)^2 \text{ and } n_2\sigma_2^2 = \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2$$

The variance of the combined series is

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2$$

or, $N\sigma^2 = \sum_{k=1}^{N}(x_k - \bar{x})^2 = \sum_{k=1}^{n_1}(x_{1i} - \bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})^2$

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

$= \Sigma\{(x_{1i} - \bar{x}_1)^2 + (\bar{x}_1 - \bar{x})^2 + 2(x_{1i} - \bar{x}_1)(\bar{x}_1 - \bar{x})\}$

$= \Sigma(\bar{x}_{1i} - \bar{x}_1)^2 + \Sigma(\bar{x}_{1i} - \bar{x})^2 + 2(\bar{x}_1 - \bar{x})\,\Sigma(\bar{x}_{1i} - \bar{x}_1)$

$= n_1\sigma_1^2 + \Sigma d_1^2 + d_1\,\Sigma(\bar{x}_{1i} - \bar{x}_1)$; Putting $d_1 = \bar{x}_1 - \bar{x}$

$= n_1\sigma_1^2 + n_1 d_1^2 + 0$; [Since, $\Sigma(x_{1i} - \bar{x}_1) = 0$]

$= n_1(\sigma_1^2 + d_1^2)$

Similarly, we get, $\sum_{j=1}^{n_2}(x_{2j} - \bar{x})^2 = n_2(\sigma_2^2 + d_2^2)$

where, $d_2 = \bar{x}_2 - \bar{x}$

$\therefore N\sigma^2 = n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)$

$\Rightarrow \sigma^2 = \dfrac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{N}$ ..... (i)

$= \dfrac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}$

$\therefore \sigma = \sqrt{\dfrac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}}$

Alternative Way :

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

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

Putting the values of d_1 and d_2 in (1) we get after simplification

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2} \quad \ldots\ldots\ldots\ldots \text{(ii)}$$

Using (i) above, the relationship can be generalized for k sets to

$$\sigma^2 = \frac{(n_1\sigma_1^2 + n_2\sigma_2^2 + \ldots\ldots + n_k\sigma_k^2) + (n_1 d_1^2 + n_2 d_2^2 + \ldots\ldots + n_k d_k^2)}{N}$$

$$= \frac{\Sigma n_i \sigma_i^2 + \Sigma n_i d_i^2}{\Sigma n_i}$$

i.e., $$\sigma = \sqrt{\frac{\Sigma n_i \sigma_i^2 + \Sigma n_i d_i^2}{\Sigma n_i}}$$

where n_i is the size of the ith set, $\bar{x}_i$ and σ_i^2 are the mean an variance respectively of the ith set, $d_i = \bar{x}_i - \bar{x}$, and σ^2 is the varian of the combined set.

Example 4.2 :

The frequency distribution of the weight of tomato (Examp 2.2) is reproduced below :

Weights:	50-60	60-70	70-80	80-90	90-100	100-110	110-
No. of tomato :	5	9	13	20	19	9	5

Calculate standard deviation by direct method and indir method.

Solution :

Direct Method :

Class interval	frequency f_i	Mid value of class x_i	$f_i x_i$	$f_i x_i^2$
50-60	5	55	275	15125
60-70	9	65	585	38025
70-80	13	75	975	73125
80-90	20	85	1700	144500
90-100	19	95	1805	171475
100-110	9	105	945	99225
110-120	5	115	575	66125
Total	N=80		6860	607600

Standard deviation $\sigma = \sqrt{\frac{1}{N}\left\{\Sigma f_i x_i^2 - \frac{(\Sigma f_i x_i)^2}{N}\right\}}$

$$= \sqrt{\frac{1}{80}\left\{607600 - \frac{(6860)^2}{80}\right\}} = \sqrt{\frac{19355}{80}}$$

$$= 15.554$$

Indirect Method :

[We change the origin to x = 85 and scale by dividing by 10]

Class interval	Mid value of class x_i	frequency f_i	$u_i = \frac{x_i - 85}{10}$	$f_i u_i$	$f_i u_i^2$
50-60	55	5	-3	-15	45
60-70	65	9	-2	-18	36
70-80	75	13	-1	-13	13
80-90	85	20	0	0	0
90-100	95	19	1	19	19
100-110	105	9	2	18	36
110-120	115	5	3	15	45
Total		N=80		6	194

$$\sigma_u = \sqrt{\frac{1}{N}\left\{\Sigma f_i u_i^2 - \frac{(\Sigma f_i u_i)^2}{N}\right\}}$$

$$= \sqrt{\frac{1}{80}\left\{(194) - \frac{(6)^2}{80}\right\}} = \sqrt{\frac{1}{80}(193.55)} \quad = 1.5554$$

$$\therefore \sigma_x = h\sigma_u = 10 \times 1.554 \quad = 15.554$$

[Note : The second method is generally known as the short-c method. But at the present age of electronic calculator it is no mor short-cut method, rather it is more lengthy and time consuming. Th is why, the method is termed here as an indirect method. Howev the method is sometimes useful when the observations of t distributions are large.]

Example 4.3 :

A student while calculating mean and standard deviation of observations obtained mean as 68 and standard deviation as 8. At t time of checking it was fond that he copied 96 instead of 69. Wh would be the actual values of mean and standard deviation ?

Solution : Here, $n = 20$, $\bar{x} = 68$ and $\sigma = 8$

We know, $\bar{x} = \frac{\Sigma x_i}{n} \Rightarrow \Sigma x_i = n\bar{x} = 20 \times 68 = 1360$

Since the student copied 69 instead of 96, the actual sum of observations is

$$\Sigma x_i = 1360 - 96 + 69 = 1333$$

$\therefore$ Actual mean, $\bar{x} = \frac{1333}{20} = 66.65$

Again we know, $\sigma^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$

$$\Rightarrow \Sigma x_i^2 = n(\sigma^2 + \bar{x}^2) = 20\,(8^2 + 68^2) = 93760$$

But actual $\Sigma x_i^2 = 93760 - 96^2 + 69^2 = 89305$

$\therefore$ Actual standard deviation is

$$\sigma = \sqrt{\frac{89305}{20} - (66.65)^2} = \sqrt{4465.25 - 4442.2225} = 4.80 \text{ (app.)}$$

Example 4.4:

Two sets of data having 200 and 250 observations have means 25 and 15 respectively and standard deviations 3 and 4 respectively. If the two sets are combined together what will be the mean and standard deviation of the combined set?

Solution : Given that,

$$n_1 = 200 \ \bar{x}_1 = 25, \sigma_1 = 3 \text{ and}$$

$$n_2 = 250 \ \bar{x}_2 = 15, \sigma_2 = 4$$

Let, mean and standard deviation of the combined set are $\bar{x}$ and σ respectively.

We know, the combined mean for two sets of observation is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{200 \times 25 + 250 \times 15}{200 + 250} = \frac{8750}{450} = 19.44$$

Again the combined standard deviation for two sets of observation is

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2}$$

$$= \sqrt{\frac{200 \times 3^2 + 250 \times 4^2}{200 + 250} + \frac{200 \times 250}{(200 + 250)^2}(25 - 15)^2}$$

$$= \sqrt{\frac{5800}{450} + \frac{5000000}{202500}} = \sqrt{12.89 + 24.69} = 6.13$$

4.2. Relative Measures of Dispersion :

- **Co-efficient of Range :** When the range is divided by the sum of highest and lowest items of the data and expressed in percentage we get the *coefficient of range* (CR).

$$\text{Thus, } CR = \frac{x_m - x_l}{x_m + x_l} \times 100\%$$

where x_m = the highest value of the data

x_l = the lowest value of the data

- **Coefficient of Quartile Deviation :** When the difference of Q_3 and Q_1 is divided by their sum and expressed in percentage, we get the *coefficient of quartile deviation* (C.Q.D).

$$\text{Thus, } CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100\%$$

where Q_3 and Q_1 are the upper and lower quartiles respectively.

- **Co-efficient of Mean Deviation :**

$$\text{CMD based upon mean, } CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}} \times 100\%$$

$$\text{CMD based upon median, } CMD(M_e) = \frac{MD(M_e)}{M_e} \times 100\%$$

$$\text{CMD based upon mode, } CMD(M_o) = \frac{MD(M_o)}{M_o} \times 100\%$$

- **Coefficient of Variation :**

Coefficient of variation of a set of data is the ratio of the standard deviation to mean expressed as percentage.

$$\text{Thus, } C.V = \frac{\sigma_x}{\bar{x}} \times 100\%$$

[Note : For comparing the variability of two series, we calculate the C.V. for each series. The series having greater C.V. is said to be more variable (unstable) than the other and the series having smaller C.V. is said to be more consistent (stable/ homogeneous) than the other. Thus C.V. is of the great practical significance and is the best measure for comparing the variability of two or more series.]

4.3. Moments :

Moments are constant which are used to determine some characteristics (e.g., nature, shape etc.) of frequency distributions. Moments about the mean are called the *central moments* and those about arbitrary value (other than mean) are known as *raw moments*.

If $x_1, x_2, \ldots\ldots, x_n$ occur with frequencies $f_1, f_2, \ldots\ldots, f_n$, respectively, then the rth central moment given by ;

$$\mu_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{N} \text{ ; where } N = \Sigma f_i \text{ ; } r = 1, 2, 3, 4, \ldots\ldots \text{ etc.}$$

In particular : $\mu_o = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^o$; when $r = 0$

$$= = \frac{1}{N}\Sigma f_i = \frac{1}{N}(N) = 1$$

1st central moment, $\mu_1 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^1$; when $r = 1$

$$\mu_1 = 0, \quad \text{Since } \Sigma f_i (x_i - \bar{x}) = 0$$

[μ_1 for any distribution is zero]

2nd central moment, $\mu_2 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^2 = \sigma^2$; when $r = 2$

[2nd central moment μ_2 is the variance]

3rd central moment, $\mu_3 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^3$; when $r = 3$

4th central moment, $\mu_4 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^4$; when $r = 4$ etc.

Raw Moment :

The rth raw moment about any arbitrary value 'a' is defined as

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a)^r \; ; \; a \neq \bar{x}$$

rth raw moment about the origin (a = 0) is $\mu'_r = \frac{1}{N}\Sigma f_i x_i^r$

when, r = 1, $\mu'_1 = \frac{1}{N}\Sigma f_i x_i = \bar{x}$

[First raw moment μ'_1 is the arithmetic mean]

$$\mu'_2 = \frac{1}{N}\Sigma f_i x_i^2 \; , \; \mu'_3 = \frac{1}{N}\Sigma f_i x_i^3 \; , \; \mu'_4 = \frac{1}{N}\Sigma f_i x_i^4 \quad \text{etc.}$$

Relation Between Central Moments and Raw Moments :

(rth central moment in terms of raw moments)

$$\mu_r = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^r$$

$$= \frac{1}{N}\Sigma f_i \left\{ x_i^r - \binom{r}{1} x_i^{r-1}(\bar{x}) + \binom{r}{2} x_i^{r-2}(\bar{x})^2 - \binom{r}{3} x_i^{r-3}(\bar{x})^3 \right.$$

$$\left. + \ldots\ldots \mp (-1)^{r-1}\binom{r}{r-1} x_i (\bar{x})^{r-1} + (-1)^r (\bar{x})^r \right\}$$

$$= \frac{1}{N}\Sigma f_i x_i^r - \binom{r}{1}\frac{1}{N}\Sigma f_i x_i^{r-1}(\bar{x}) + \binom{r}{2}\frac{1}{N}\Sigma f_i x_i^{r-2}(\bar{x})^2 + \binom{r}{3}\frac{1}{N}\Sigma f_i x_i^{r-3}(\bar{x})^3$$

$$+ \ldots\ldots \mp (-1)^{r-1}\binom{r}{r-1}\frac{1}{N} f_i x_i (\bar{x})^{r-1} + (-1)^r \frac{1}{N} f_i (\bar{x})^r$$

Putting $\bar{x} = \mu'_1$, 1st raw moment about the origin, we get

$$\mu_r = \mu'_r - \binom{r}{1}\mu'_{r-1}\mu'_1 + \binom{r}{2}\mu'_{r-2}(\mu'_1)^2 - \binom{r}{3}\mu'_{r-3}(\mu'_1)^3$$

$$+ \ldots\ldots \mp (-1)^{r-1} r \mu'_1 (\mu'_1)^{r-1} + (-1)^r (\mu'_1)^r$$

When, r = 2

$$\mu_2 = \mu'_2 - \binom{2}{1}\mu'_{2-1}\mu'_1 + \binom{2}{2}\mu'_{2-2}(\mu'_1)^2 = \mu'_2 - 2\mu'_1\mu'_1 + \mu'_0(\mu'_1)^2$$

$$= \mu'_2 - 2(\mu'_1)^2 + (\mu'_1)^2 \qquad \left[\text{Since } \mu'_0 = \frac{1}{N}\Sigma f_i x_i^0 = 1\right]$$

$$= \mu'_2 - \mu'^2_1$$

- When, r = 3

$$\mu_3 = \mu_3' - \binom{3}{1}\mu_{3-1}'\,\mu_1' + \binom{3}{2}\mu_{3-2}'\,(\mu_1')^2 - \binom{3}{3}\mu_{3-3}'\,(\mu_1')^3$$

$$= \mu_3' - 3\mu_2'\,\mu_1' + 3\mu_1'\,\mu_1'^2 - \mu_0'\,\mu_1'^3$$

$$= \mu_3' - 3\mu_2'\,\mu_1' + 3\mu_1'^3 - \mu_1'^3$$

$$= \mu_3' - 3\mu_2'\,\mu_1' + 2\mu_1'^3$$

- When, r = 4

$$\mu_4 = \mu_4' - \binom{4}{1}\mu_{4-1}'\,\mu_1' + \binom{4}{2}\mu_{4-2}'\,\mu_1'^2 - \binom{4}{3}\mu_{4-3}'\,\mu_1'^3 + \binom{4}{4}\mu_{4-4}'\,\mu_1'^4$$

$$= \mu_4' - 4\mu_3'\,\mu_1' + 6\mu_2'\,\mu_1'^2 - 4\mu_1'\,\mu_1'^3 + \mu_0'\,\mu_1'^4$$

$$= \mu_4' - 4\,\mu_3'\,\mu_1' + 6\mu_2'\,\mu_1'^2 - 3\mu_1'^4$$

Moments are Independent of Change of Origin but not of Scale.

Proof. :

Let, $x_1, x_2, \ldots\ldots\ldots, x_n$ be the mid-values of the classes of a frequency distribution and let $f_1, f_2, \ldots\ldots, f_n$ be their corresponding frequencies.

Now rth central moment is

$$\mu_{r(x)} = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^r$$

We change the origin and scale of x such that

$$u_i = \frac{x_i - a}{h} \implies \bar{u} = \frac{\bar{x} - a}{h}$$

Now; for new variate u ; we have

$$\mu_{r(u)} = \frac{1}{N}\Sigma f_i (u_i - \bar{u})^r$$

$$= \frac{1}{N}\Sigma f_i \left\{\frac{x_i - a}{h} - \frac{\bar{x} - a}{h}\right\}^r$$

$$= \frac{1}{N}\Sigma f_i \left(\frac{x_i - a - \bar{x} + a}{h}\right)^r = \frac{1}{h^r}\frac{1}{N}\Sigma f_i (x_i - \bar{x})^r$$

$$\therefore \mu_{r(u)} = \frac{1}{h^r} \mu_{r(x)}$$

$$\therefore \mu_{r(x)} = h^r \mu_{r(u)}$$

Hence, moments are independent of original but dependent on scale. Proved.

- **Sheppard's Correction for Moments :**

In calculating the moments of a grouped frequency distribution we assume that all the values within a class interval refer to mid-value of the class interval. If the distribution is symmetrical or moderately asymmetrical and the class intervals are small (i.e. greater that $\frac{1}{20}$th Range), this assumption is approximately true.

Generally, this assumption is not always true, some error, called grouping error creeps into the calculation of the moments.

W.F. Sheppard proposed that if

(i) the frequency distribution is continuous and

(ii) the frequency tapers off to zero in both ends of the class interval the effect due to grouping at the mid-point of the intervals can be corrected by the following formulae, known as Sheppard's Corrections :

$$\mu_2 \text{ (correlated)} = \mu_2 - \frac{h^2}{12}$$

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

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4$$

where h is the length of class interval.

Example 4.5 : The wages per hour of 100 farm labours are given below :

Wages (Taka)	: 0-5	5-10	10-15	15-20	20-25
No. of labours	: 10	15	40	25	10

Compute first four central moments (use Sheppard's-correction for the 2nd and 4th central moments) :

Solution :

Wages (Tk.)	No. of labours f_i	Mid value x_i	$u_i = \frac{x_i - 12.5}{5}$	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
0-5	10	2.5	-2	-20	40	-80	160
5-10	15	7.5	-1	-15	15	-15	15
10-15	40	12.5	0	0	0	0	0
15-20	25	17.5	1	25	25	25	25
20-25	10	22.5	2	20	40	80	160
Total	100			10	120	10	360

$$\mu'_{1(u)} = \frac{1}{N}\Sigma f_i u_i = \frac{1}{100} \times 10 = 1.0$$

$$\mu'_{2(u)} = \frac{1}{N}\Sigma f_i u_i^2 = \frac{1}{100} \times 120 = 1.2$$

$$\mu'_{3(u)} = \frac{1}{N}\Sigma f_i u_i^3 = \frac{1}{100} \times 10 = 0.1$$

$$\mu'_{4(u)} = \frac{1}{N}\Sigma f_i u_i^4 = \frac{1}{100} \times 360 = 3.6$$

Now, $$\mu_{2(u)} = \mu'_{2(u)} - \left\{\mu'_{1(u)}\right\}^2 = 1.2 - (0.1)^2 = 1.19$$

$$\mu_{3(u)} = \mu'_{3(u)} - 3\mu'_{2(u)}\mu'_{1(u)} + 2\left\{\mu'_{1(u)}\right\}^3$$

$$= 0.1 - 3(1.2)(0.1) + 2\{0.1\}^3 = -0.258$$

$$\mu_{4(u)} = \mu'_{4(u)} - 3\mu'_{3(u)}\mu'_{1(u)} + 6\mu'_{2(u)}\left\{\mu'_{1(u)}\right\}^2 - 3\left\{\mu'_{1(u)}\right\}^4$$

$$= 3.6 - 4(0.1)(0.1) + 6(1.2)(0.1)^2 - 3(0.1)^4$$

$$= 3.6 - 0.04 + 0.072 - 0.003 = 3.6317$$

First Four Central Moments of the Original Variable:

$$\mu_{1(x)} = 0$$

$$\mu_{2(x)} = h^2\mu_{2(x)} = (5)^2(1.19) = 29.75$$

$$\mu_{3(x)} = h^3\mu_{3(x)} = (5)^3(-0.258) = -32.25$$

$$\mu_{4(x)} = h^4\mu_{4(x)} = (5)^4(3.6317) = 2269.8125$$

Application of Sheppard's correction for moments:

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12} = 29.75 - \frac{(5)^2}{12} = 27.667$$

$$\mu_3 \text{ (corrected)} = \mu_3 = -32.25$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4$$

$$= 2269.8125 - \frac{(5)^2}{2}(29.75) + \frac{7}{240}(5)^4$$

$$= 1916.1667$$

4.4. Skewness :

Skewness means lack of symmetry. For an asymmetric distribution it is the departure from symmetry.

Symmetrical Distribution : A distribution is said to be symmetrical if the frequencies are symmetrically distributed about the mean. For symmetrical distributions the values equi-distant from mean have equal frequency. For example, the following distribution is symmetrical about its mean 4.

x :	0	1	2	3	4	5	6	7	8
f :	12	14	16	18	20	18	16	14	12

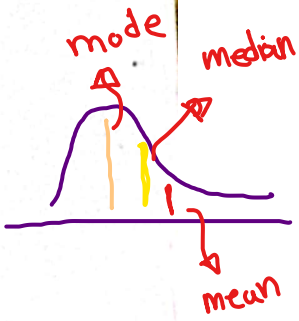
Again for symmetrical distribution mean = mode = median.

A distribution is said to be skewed if .

i) Mean, median and mode fall at different points.

ii) Q_1 and Q_3 are not equidistant from median and

iii) The curve drawn with the help of the given data is not symmetrical but elongated more to one side,

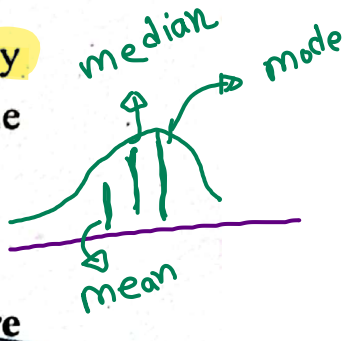


Skewness may be positive or negative. Skewness is said to be positive if the frequency curve is more elongated to the right side. In this case mean of the distribution lies at the right of (or greater than) the mode.

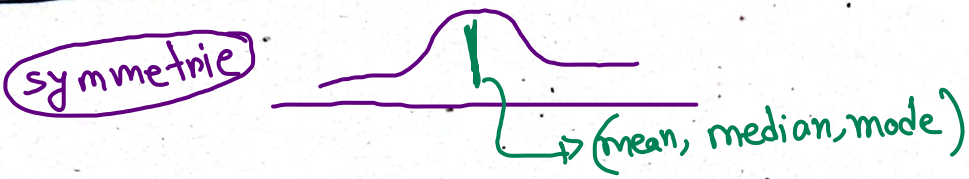
i.e. $\bar{x} > M_e > M_o$.

On the otherhand, the skewness is negative if the frequency curve is more elongated to the left side. In this case mean of the distribution lies at the left of (or less than) the mode.

i.e. $M_o > M_e > \bar{x}$



For symmetrical distributions the mean, median and mode are same.



Proof*:

Let us consider the following symmetrical continuous frequency distribution with equal class interval ($x_1 < x_2 < \ldots\ldots < x_{n+1}$):

Table 4.1: Frequency distribution.

Class interval	Mid value y_i	Frequency f_i	Cumulative frequency F_i
$x_1 - x_2$	y_1	f_1	F_1
$x_2 - x_3$	y_2	f_2	F_2
$x_3 - x_4$	y_3	f_3	F_3
: :	:	:	:
$x_{k-1} - x_k$	y_{k-1}	f_{k-1}	F_{k-1}
$x_k - x_{k+1}$	y_k	f_k	F_k
$x_{k+1} - x_{k+2}$	y_{k+1}	f_{k+1}	F_{k+1}
: :	:	:	:
$x_{n-1} - x_n$	y_{n-1}	f_{n-1}	F_{n-1}
$x_n - x_{n+1}$	y_n	f_n	F_n

*Adopted with minor modification from an unpublished article by M. Amirul Islam providing a theoretical proof.]

Since the distribution is symmetrical, we will have $f_1 = f_{n-1}$, $f_2 = f_{n-2}$, ..., $f_{k-1} = f_{k+1}$ and f_k will be the highest frequency. Let us consider that h be the width of each class interval.

From the traditional formula of mode we get,

$$M_o = L_o + \frac{f_o - f_1}{2f_o - f_1 - f_2} \times h$$

$$= x_k + \frac{h(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$$

[Putting $L_o = x_k$, $f_o = f_k$, $f_1 = f_{k-1}$ and $f_2 = f_{k+1}$]

$$= x_k + \frac{h(f_k - f_{k-1})}{2f_k - 2f_{k-1}}; \quad [\text{since } f_{k-1} = f_{k+1}]$$

$$= x_k + \frac{h}{2}$$

Since the distribution is symmetrical we get,

$$f_1 + f_2 + \ldots\ldots + f_{k-1} + \tfrac{1}{2}f_k = \tfrac{1}{2}f_k + f_{k+1} + f_{k+2} + \ldots\ldots + f_n$$

$$\Rightarrow \frac{N}{2} = f_1 + f_2 + \ldots\ldots + f_{k-1} + \frac{1}{2}f_k, \text{ where } N = \sum_{i=1}^{n} f_i$$

$$\Rightarrow \frac{N}{2} = F_{k-1} + \frac{1}{2}f_k \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Again, from the traditional formula of median we get,

$$M_e = L_m + \frac{\frac{N}{2} - F'_m}{f_m} \times h$$

$$= x_k + \frac{\frac{N}{2} - F_{k-1}}{f_k} \times h$$

[putting $L_m = x_k$, $f_m = f_k$ and $F'_m = F_{k-1}$]

$$= x_k + \frac{(F_{k-1} + \tfrac{1}{2}f_k) - F_{k-1}}{f_k} \times h$$

$$= x_k + \frac{\frac{1}{2}f_k}{f_k} \times h$$

$$= x_k + \frac{h}{2}$$

We know, for a frequency distribution

Arithmetic mean, $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i y_i$ ..... (2)

For a symmetric distribution

$$\sum_{i=1}^{n} f_i y_i = f_1y_1 + f_2y_2 + \ldots + f_{k-1}y_{k-1} + f_ky_k + f_{k+1}y_{k+1} + \ldots + f_ny_n$$

$$= (f_1y_1 + f_ny_n) + (f_2y_2 + f_{n-1}y_{n-1}) + (f_{k-1}y_{k-1} + f_{k+1}y_{k+1}) + f_ky_k$$

$$= f_1(y_1 + y_2) + f_2(y_2 + y_{n-1}) + \ldots + f_{k-1}(y_{k-1} + y_{k+1}) + f_ky_k \ldots (3)$$

$$[f_1 = f_n, f_2 = f_{n-1}, \ldots, f_{k-1} = f_{k+1}]$$

As the distribution is symmetrical we will also have

$(y_1 + y_n) = (y_2 + y_{n-1}) = \ldots = (y_{k-1} + y_{k+1}) = 2y_k$

Putting these values in equation (3) we get,

$$\sum_{i=1}^{n} f_i y_i = 2y_k (f_1 + f_2 + \ldots + f_{k-1}) + f_ky_k$$

$$= 2y_k (f_1 + f_2 + \ldots + f_{k-1} + \tfrac{1}{2}f_k)$$

$$= 2y_k \frac{N}{2} \qquad \text{[From equation 1]}$$

$$= y_k \cdot N$$

Putting this value is in equation (2) we get,

Arithmetic mean $= \frac{1}{N} y_k N = y_k$

Since y_k is the mid value of the class $(x_k - x_{k+1})$ having class interval h,

We get, $y_k = x_k + \frac{h}{2}$

$\Rightarrow \bar{x} = x_k + \frac{h}{2}$

Hence, Arithmetic Mean = Median = Mode. Proved.

Position of Mean, Median and Mode :

The position of arithmetic mean, median and mode of symmetrical frequency distribution is shown in Figure 4.1.

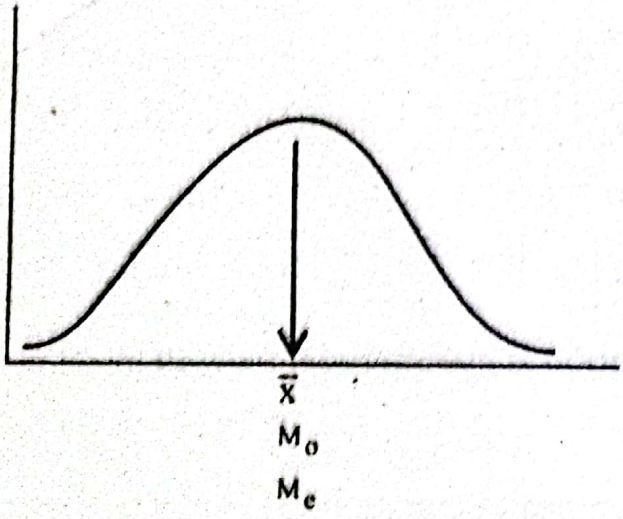


Fig. 4.1.

For distributions of moderate skewness, there is an empiri relationship among the mean, median and mode that,

Mean - Mode = 3(Mean - M_e)

or, $\bar{x} - M_o = 3(\bar{x} - M_e)$

The position of arithmetic mean, median and mode moderately asymmetrical distributions are shown in Fig. 4.2 and F 4.3.

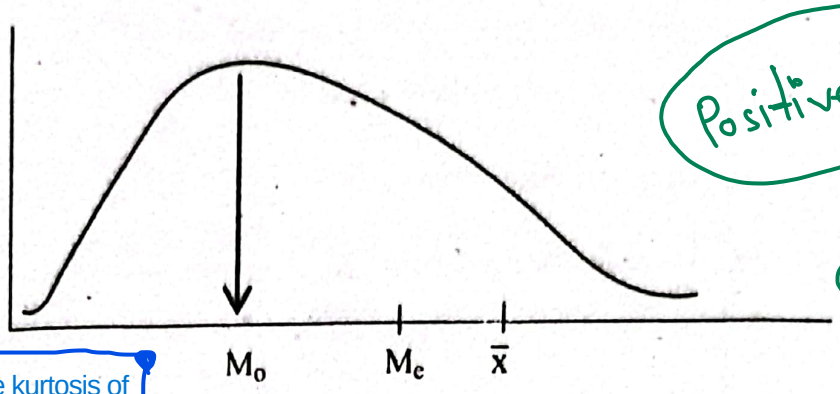


Fig. 4.2

Positive skwed or right skwed

Beta(2) This measures the kurtosis of a distribution, which tells how "peaked" or "flat" the distribution is compared to a normal distribution.

For a normal distribution, B2 = 3, the distribution is normal peaked (mesokurtic),

B2>3, the distribution is more peaked (leptokurtic),

B2<3, it is flatter (platykurtic

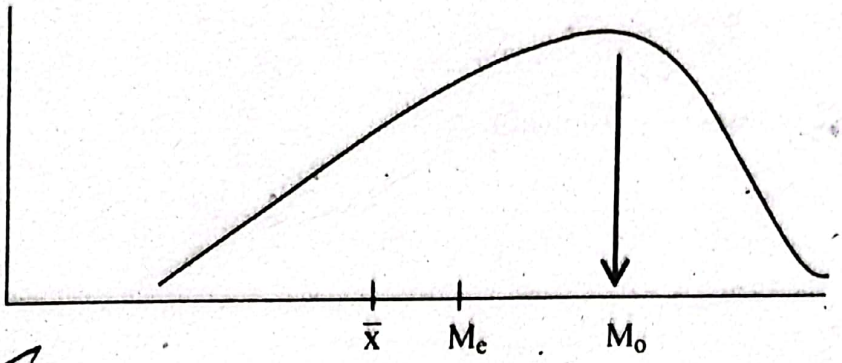


Fig. 4.3

left-skwed negative skwed

Karl Pearson's β and γ Co-efficient :

Karl Pearson defined the following co-efficients, based upon first four central moments :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

B1 = This measures the skewness of a distribution. Skewness refers to the asymmetry of the distribution about its mean.
where,
Mu(2) is the second central moment (variance), Mu(3) is the third central moment (which measures skewness). If
b1 = 0. the distribution is symmetric.

$$\gamma_1 = \pm\sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3$$

A positive value of ,
gamma_1 indicates right skewness, and a negative value of gamma_1 indicates left skewness.

Measures of Skewness :

We may compare the nature, shape and size of two or more frequency distributions with the help of measures of skewness. The difference between mean and mode is considered as a measure of skewness. If $\bar{x} > M_e$ the skewness is said to be positive and if $\bar{x} < M_e$, the skewness is said to be negative. Skewness of distributions

having different units of measurement cannot be compared with th help of absolute measures of skewness. That is why, relati measures of skewness are widely used.

Relative Measures of Skewness :

(1) Karl Pearson's Formula. $S_k = \frac{\text{Mean} - \text{Mode}}{\text{s.d.}} = \frac{\bar{x} - M_o}{\sigma}$

In case it is not possible to find the mode or if a distribution ha more than one mode, the following formula is used to measu skewness : $S_k = \frac{3(\text{Mean} - \text{Median})}{\text{s.d.}} = \frac{3(\bar{x} - M_e)}{\sigma}$

(2) Bowley's formula

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

$$= \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1}$$

where Q_1, Q_2 and Q_3 are the 1st, 2nd and 3rd quartiles respectively.

(3) Keley's formula :

$$S_k = \frac{D_9 + D_1 - 2M_e}{D_9 - D_1} \text{ or, } S_k = \frac{P_{99} + P_1 - 2M_e}{P_{99} - P_1}$$

(4) Co-efficient of skewness based upon moments.

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

As both β_1 and β_2 are always non-negative, the above form cannot indicate as to whether the skewness is positive or negative such case the nature of the distribution will depend upon the value

μ_3. If μ_3 is positive, the skewness is considered to be positive and if μ_3 is negative the skewness is also treated to be negative.

4.5. Kurtosis:

Like skewness, *kurtosis* is also an important shape characteristic of frequency distribution. Two distributions may be both symmetrical, they may have the same variability as measured by standard deviation, they may be relatively more or less flat topped compared to normal curve (Discussed in chapter VII). This relative flatness of the top or the degree of peakedness is called *kurtosis* and is measured by β_2. For normal distribution, $\beta_2 = 3$. Hence the quantity β_2-3 is known as excess of kurtosis or simply kurtosis. On the basis of kurtosis, frequency curves are divided into the following three categories :

1) Leptokurtic ; a curve having a high peak.
2) Platykurtic ; a curve which is flat topped
3) Mesokurtic ; a curve which is neither too peaked nor too flat-topped.

For formal distribution, $\beta_2 = 3$ and $\gamma_2 = 0$. Kurtosis is measured by $\gamma_2 = \beta_2 - 3$.

If a distribution has

(i) $\beta_2 > 3$, it is called leptokurtic
(ii) $\beta_2 < 3$, it is called platykurtic
(iii) $\beta_2 = 3$, it is called mesokurtic

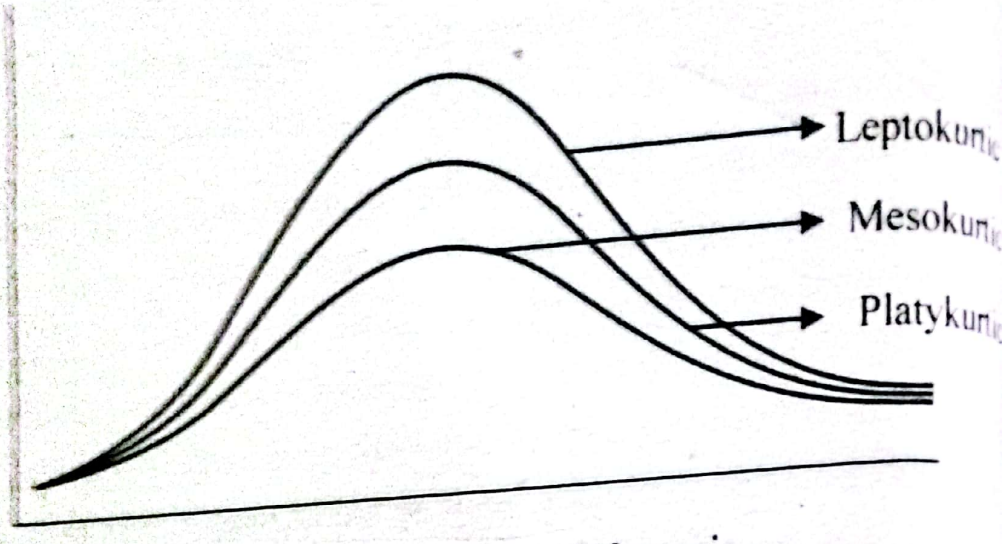


Fig. 4.4: Different types of Kurtosis

Example 4.25 : A distribution of short term agricultural credit disbursement from 10 branches of a bank is given below -

Amount of credit : (Lac Taka)	0-5,	5-10,	10-15,	15-20,	20-25
No. of branches :	1	2	4	2	1

Calculate first four central moments, co-efficients of skewness and kurtosis and thus comment on the shape and nature of the distribution.

Solution :

Amount of credit (lac Tk.)	No. of branches f_i	Mid value x_i	f_ix_i	$x_i - \bar{x}$	$f_i(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^3$	$f_i(x_i -$
0-5	1	2.5	2.5	-10	-10	100	-1000	100
5-10	2	7.5	15.0	-5	-10	50	-250	125
10-15	4	12.5	50.0	0	0	0	0	0
15-20	2	17.5	35.0	5	10	50	250	125
20-25	1	22.5	22.5	10	10	100	1000	100
Total	N=10		125.0	0	0	300	0	225

$$\bar{x} = \frac{1}{N}\Sigma f_i x_i = \frac{1}{10} \times 125.0 = 12.5$$

$$\mu_1 = \frac{1}{N}\Sigma f_i (x_i - \bar{x}) = \frac{1}{10} \times (0) = 0 \ (\mu_1 = 0 \text{ always})$$

$$\mu_2 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^2 = \frac{1}{10} \times (300) = 30.0$$

$$\mu_3 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^3 = \frac{1}{10} \times (0) = 0$$

ınd $$\mu_4 = \frac{1}{N}\Sigma f_i (x_i - \bar{x})^4 = \frac{1}{10} \times (22500) = 2250.0$$

Now, $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{0}{(30)^3} = 0$

$\therefore$ Coefficient of skewness $S_k = \dfrac{(\beta_2 + 3)\sqrt{\beta_1}}{2(5\beta_2 - 6\beta_1 - 9)} = 0$

Hence the distribution is symmetrical.

Again $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{2250}{(30)^2} = 2.5 < 3$

$\therefore \gamma = \beta_2 - 3 = 2.5 - 3 = -0.5.$

Since $\gamma < 0$; The curve is platykurtic.

$\therefore$ The distribution is symmetrical and platykurtic.