CHAPTER-I

DATA EXPLORATION AND REPRESENTATION

The word óStatisticsô is probably derived from the Latin word óstatusô (means a political state) or the Italian word óstatistaô or the German word óstatistikô, each of which means a ópolitical stateô. It is used in singular as well as in plural sense. In singular sense, statistics is used as a subject that deals with the principles and methods employed in collection, presentation, analysis and interpretation of data. In plural sense, statistics is considered as numerical description of quantitative information.

1.1 Statistics (Definition), Scope and Limitations:

Different persons defined Statistics in different ways. Some of the popular definitions of Statistics are given below.

According to Croxton and Cowden, ñ*Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data*ò.

Professor Horace Secrist defined Statistics as ñ*aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other*ò.

According to Sir R.A. Fisher ñ*The science of Statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data*ò.

Fisherôs definition is most exact in the sense that it covers all aspects and fields of Statistics. On the basis of above ideas, Statistics can be defined as a science which deals with collection, presentation, analysis of data and interpretation of results.

Statistics is often classified as Mathematical Statistics and Applied Statistics. The Mathematical Statistics deals with the theoretical developments, derivation of the formulae and statistical solutions of the problems. Applied Statistics on the other hand deals with the application of the statistical methods in different branches of science, art and industry.

The credit for applications of statistics in various diverse fields goes to Sir R.A. Fisher (1890-1962) known as ñFather of Statisticsò. Today, the statistics is not merely

confined to the affairs of the state for data collection but it is regarded as an important tool of all physical, social and biological sciences indispensable to research and intelligent judgement. The scope of statistics is given below:

Scope of Statistics:

i) Statistics has great significance in the field of physical and natural sciences. It is used in propounding and verifying scientific laws. Statistics is often used in agricultural and biological research for efficient planning of experiments and for interpreting experimental data.

ii) Statistics is of vital importance in economic planning. Priorities of planning are determined on the basis of the statistics related to the resource base of the country and the short-term and long-term needs of the country.

iii) Statistical techniques are used to study the various economic phenomena such as wages, price analysis, analysis of time series, demand analysis etc.

iv) Successful business executives make use of statistical techniques for studying the needs and future prospects of their products. The formulation of a production plan in advance is a must, which cannot be done in absence of the relevant details and their proper analysis, which in turn requires the services of a trained statistician.

v) In industry, the statistical tools are very helpful in the quality control and assessment. In particular, the inspection plans and control charts are of immense importance and are widely used for quality control purposes.

Limitations of Statistics:

i) Statistical methods are best applicable to quantitative data.

ii) Statistical decisions are subject to certain degree of error.

iii) Statistical laws do not deal with individual observations but with a group of observations.

iv) Statistical conclusions are true on an average.

v) Statistics is liable to be misused. The misuse of statistics may arise because of the use of statistical tools by inexperienced and untrained persons.

vi) Statistical results may lead to fallacious conclusions if quoted out of context or manipulated.

1.2 Some Basic Concepts:

Variable: A quantity that varies from individual to individual is called a variable. Height, weight, number of students in a college, number of petals in a flower, number of tillers in a plant etc. are a few examples of variables.

Discrete and Continuous Variables: A variable that takes only specific or distinguished values in a given range is known as discrete variable whereas a variable which can theoretically assume any value between two given values is called a continuous variable. For example, the number of students in a college, number of petals in a flower, number of tillers in a plant etc. are discrete variables. A continuous variable can take any value within a certain range, for example, yield of a crop, height of plants and birth rates etc. are continuous variables.

Qualitative and Quantitative Characters:

Those characteristics/attributes of individuals which cannot be measured numerically e.g. sex, blindness, honesty, colour etc. are called qualitative characters whereas the characteristics of the individuals which can be numerically measured, e.g. height, weight, age, yield etc. are called quantitative characters.

Raw Data and Array:

The collected data which have not been processed or organized numerically is known as raw data while, the raw data arranged in ascending or descending order of magnitude is called an array.

Primary and Secondary Data:

The data collected directly from the original source are called the primary data. Such data may be collected by sample surveys or through designed experiments. The data, which have already been collected by some agency and have been processed or used at least once are called secondary data. Secondary data may be collected from organizations or private agencies, government records, journals etc.

Classification of Data: It is process of arranging the data into a number of classes or groups on the basis of their resemblances and similarities. It is of four types:

i) **Geographical Classification:** The data may be classified according to geographical or locational differences such as regions, states, districts, cities etc.

for example, data on sale of automobiles in different states and districts in India in a particular year.

ii) **Chronological Classification:** Here the data are arranged according to time started from the some initial time period, For example, the data on sale of automobile in India over the last 10 years.

iii) **Qualitative Classification:** This type of classification is applicable for qualitative data and the data are classified according to some characteristic or attribute such as religion, sex, employment, national origin etc. The attributes cannot be measured but can be categorized e.g. the population of a town may be classified as follows:

Sex / **Employment Status**	**Male**	**Female**	**Total**
Employed	4600	940	**5540**
Unemployed	510	3950	**4460**
Total	**5110**	**4890**	**10,000**

iv) **Quantitative Classification:** It is for quantitative data like data on weight, height, income etc. of individuals. Here the construction of a frequency distribution is required.

Difference between Classification and Tabulation:

i) The classification is the basis of tabulation while tabulation is a mechanical function of classification.

ii) The classification divides the data into homogeneous groups and subgroups with regard to the similarity of the characteristics under study while tabulation arranges the classified data into rows and columns with regard to time, size, aim and importance of data.

iii) The classification is a technique of statistical analysis while tabulation is a technique of presenting the data.

iv) Classification facilitates the comparison between the two data sets while tabulation makes a comparison very easy through the use of ratios, percentages and coefficients etc.

Tabulation: It is the process in which the data are put in a table having different rows and columns. A table, which contains data related to one characteristic, is called a simple table. On the other hand, a table which contains data related to more than one characteristic, is called a complex table.

1.3 Frequency Distributions and Their Construction:

Frequency Distribution:

The number of observations lying in any class interval is known as the frequency of that class interval. Also the number of times an individual item is repeated in a series is called its frequency. The way in which the observations are classified and distributed in the proper class intervals is known as frequency distribution.

Relative Frequency:

It is the proportion of the number of observations belonging to a class and is obtained by dividing the frequency of that class by the total frequency.

Cumulative Frequency:

The cumulative frequency corresponding to any value or class is the number of observations less than or equal to that value or upper limit of that class. It may also be defined as the total of all frequencies up to the value or the class.

Cumulative Frequency Distribution:

It is an arrangement of data in class intervals together with their cumulative frequencies. In less than cumulative frequency distribution, the frequency of each class is added successively from the top to bottom but in more than type, the frequencies of each class are added successively from bottom to top.

Rules for Construction of Frequency Distribution:

i) The number of classes should preferably be between 5 and 15, however, there is no rigidity about it and it depends upon total frequency and the range of the data. The following formula suggested by H.A. Sturges may be used for finding approximate number of class intervals and their width.

$$h = \frac{L - S}{K}$$

where N = total frequency ; $K = 1 + 3.322 \log_{10} N$ is the number of classes. L and S are the largest and smallest observation in the data and h = class width.

ii) The class limits should be well defined so that one can place an observation in a class without any confusion.

iii) As far as possible, the class intervals should be of equal size.

iv) Counting of number of observations belonging to a class gives the frequency of the class.

Discrete (ungrouped) Frequency Distribution:

When the number of observations in the data are small, then the listing of the frequency of occurrence against the values of variable is called as discrete frequency distribution.

Example-1: The number of seeds per pod in 50 pods of a crop variety is given below: Prepare a discrete frequency distribution.

9, 2, 3, 1, 4, 5, 2, 6, 2, 3, 8, 9, 7, 6, 5, 4, 1,
3, 2, 7, 5, 4, 5, 4, 3, 8, 7, 5, 4, 3, 6, 5, 3, 4, 8, 6,
8, 5, 4, 7, 3, 4, 5, 6, 9, 8, 5, 1, 4, 5

Solution: The number of seeds can be considered as a variable (X) and the number of pods as the frequency (f).

No. of Seeds (X)	Tally Marks	No. of Pods (f)
1	III	3
2	IIII	4
3	~~IIII~~ II	7
4	~~IIII~~ IIII	9
5	~~IIII IIII~~	10
6	~~IIII~~	5
7	IIII	4
8	~~IIII~~	5
9	III	3
Total:	**N**	**50**

Grouped (Continuous) Frequency Distribution:

When the data set is very large it becomes necessary to condense the data into a suitable number of class intervals of the variable alongwith the corresponding frequencies. The following two methods of classification are used

Exclusive Method: When the data are classified in such a way that the upper limit of a class interval is the lower limit of the next class interval, then it said to be the exclusive method of classification i.e. upper limits are not included in the class interval.

Inclusive Method: When the data are classified in such a way that both lower and upper limits are included in the class interval, then it said to be inclusive method of classification.

Remarks:

1. An exclusive method should be used for the continuous data and inclusive method may be used for discrete data.
2. If the continuous data are classified according the inclusive method, there is need to find class boundaries to maintain continuity let d be the difference between the upper limit of a class and lower limit of the next class, then

   lower class boundary = lower limit - 0.5d.

   and upper class boundary = upper limit + 0.5d.

   i.e. to obtain the class boundaries, take the difference (d) between 20 and 19 which is one. Thus d/2 = 0.5 Now deduct 0.5 from the lower limits and 0.5 to the upper limits.

Table: Class intervals showing Dividend declared by 45 companies

Exclusive method	Inclusive Method	Class Boundaries	Frequency (Number of Companies)
10-20 (10 but less than 10)	10-19	9.5-19.5	7
20-30 (20 but less than 30)	20-29	19.5-29.5	13
30-40 (30 but less than 40)	30-39	29.5-39.5	15
40-50 (40 but less than 50)	40-49	39.5-49.5	10

Example-2: The marks obtained by 30 students are given below. Classify the data using Sturges rule

30, 32, 45, 52, 47, 52, 58, 63, 59, 75, 49, 55, 77,

28, 26, 33, 47, 45, 59, 73, 75, 65, 55, 68, 67, 79, 35, 39, 68, 75

Let us find the suitable number of class intervals with the help of Sturges rule

No. of classes (K) = $1 + 3.322 \log_{10} N = 1 + 3.322 \log_{10} 30 = 1 + (3.322 \times 1.477)$

$= 1 + 4.91 = 5.91 \simeq 6$

$$\text{Class width (h)} = \frac{L - S}{K} = \frac{79 - 26}{6} = \frac{53}{6} = 8.8 \simeq 9$$

Thus number of classes will be 5.91 (=6) and size of class interval will be 9

We take 10 as the size of the class interval. Since the minimum value is 26, therefore, the first class interval is taken as 25-35.

Marks Obtained (Class Interval)	Tally Marks	No. of Students (f)
25-35	~~HH~~ I	5
35-45	II	2
45-55	~~HH~~ II	7
55-65	~~HH~~ I	6
65-75	~~HH~~	5
75-85	~~HH~~	5
	N	**30**

1.4 Measures of Central Tendency:

Different observations have a tendency to concentrate around a central point in a data series which is often called central tendency. A central tendency or an average can also be defined as a single value within the range of the data that represents all the values in the data series. Since an average is somewhere within the range of the data, it is sometimes called a measure of central value or location of a distribution.

Arithmetic mean, geometric mean, harmonic mean, median and mode are the popular measures of central tendency. The arithmetic mean, geometric mean and harmonic mean are known as mathematical averages while median and mode are positional averages.

Properties of a Good Measure of Central Tendency:

i) It should be rigidly defined, based on all the observations and easy to calculate and understand.

ii) It should be suitable for further mathematical treatment.

iii) It should be least affected by extreme values and fluctuations in sampling.

Arithmetic Mean:

Arithmetic mean of a group of n observations $x_1, x_2, ..., x_n$ is defined as

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum x_i$$

If the observations $x_1, x_2, ..., x_n$ form a discrete frequency distribution with respective frequencies $f_1, f_2, \cdots\cdots, f_n$ then arithmetic mean is given by:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + ... + f_nx_n}{N} = \frac{1}{N}\sum f_i x_i \text{ where } N = \Sigma f_i$$

For grouped frequency distribution, mid values are determined for the various class intervals and then arithmetic mean is computed as in case of a discrete frequency distribution.

Properties of Arithmetic Mean:

i) Arithmetic mean is rigidly defined and based on all the observations.

ii) It is suitable for further mathematical treatment and is least affected by fluctuations in samplings.

iii) Sum of deviations of the given values from their arithmetic mean is always zero.

iv) Sum of the squares of deviations of the given values from their arithmetic mean is always minimum.

Demerits of Arithmetic Mean:

i) Arithmetic mean cannot be used as a suitable measure of central tendency if we are dealing with qualitative characteristics or in case of extremely asymmetrical distributions.

ii) Arithmetic mean is likely to be affected by extreme values.

iii) Arithmetic mean cannot be calculated in case of open-ended classes.

Pooled or Combined Arithmetic Mean: If $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ are the arithmetic means of k groups, based on $n_1, n_2, \cdots\cdots, n_k$ observations respectively then combined mean of all ($n_1 + n_2 + ..$é n_k) observations of the k groups taken together is:

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

Weighted Mean: When different values have unequal weightage or importance or contributions then instead of simple mean the weighted mean is used. Let $x_1, x_2, \cdots\cdots, x_k$ be k values with weights $w_1, w_2, \cdots\cdots, w_k$ respectively then weighted mean is

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + ... + w_kx_k}{\sum w_i} = \frac{\sum w_i x_i}{\sum w_i}$$

Example-3: Find over all grade point average (OGPA) in a semester from the following:

Course	**Cr. Hrs. (w_i)**	**Grade point (x_i)**	**$w_i x_i$**
Stat-101	2+1	6.3	18.9
Comp-101	1+1	7.0	14.0
Math-101	3+1	7.5	30.0
Total:	**6+3**		**62.9**

OGPA is $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{62.9}{9} = 6.99$

Geometric Mean: Geometric mean is more appropriate if the observations are measured as ratios, proportions, growth rates or percentages. When the growth rates or increase in production etc. are given for a number of years or periods then geometric mean should be used as a measure of central tendency. If $x_1, x_2, \cdots, x_n$ be the n positive observations then $G = (x_1 x_2 \ldots é \ x_n)^{1/n}$ or $\log G = \frac{1}{n}\sum \log x_i$ For a frequency distribution, the geometric mean is defined as

$$G = (x_1^{f1}.x_2^{f2} é \ .. \ x_n^{fn})^{1/N} \quad \text{or} \quad \log G = \frac{1}{N}\sum f_i \log x_i \text{ ,where } N = \Sigma f_i$$

Properties of Geometric Mean:

i) It is rigidly defined and is based on all the observations.

ii) It is suitable for further mathematical treatment. If n_1 and n_2 are the sizes of two data series with G_1 and G_2 as their geometric means respectively, then geometric mean of combined series G is given by:

$$\text{Log } G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

iii) It is not affected much by fluctuations in sampling.

Demerits:

i) Geometric mean is a bit difficult to understand and to calculate for a non-mathematician.

ii) It cannot be calculated when any value is zero or negative and gives an absurd value if computed in case of even number of negative observations

iii) Like arithmetic mean it is also affected by the extreme values but to a lesser extent.

Harmonic Mean: The reciprocal of the arithmetic mean of reciprocals of non-zero values of a variable is called harmonic mean (H). Harmonic mean is a suitable average when observations are given in terms of speed rates and time.

If $x_1, x_2, \cdots\cdots, x_n$ are n observed values, the harmonic mean is given by:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}}$$

When observations are given in the form of a frequency distribution, then harmonic mean is given by:

$$H = \frac{N}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots\ldots + \frac{f_n}{x_n}} = \frac{N}{\sum \frac{f_i}{x_i}} \text{ where } N = \sum f_i$$

Harmonic mean for a grouped frequency distribution can be obtained by first calculating the class marks.

Properties of Harmonic Mean:

i) Harmonic mean is rigidly defined and is based on all observations.

ii) It is suitable for further mathematical treatment.

iii) Like geometric mean it is not affected much by fluctuation of sampling.

Demerits:

i) Harmonic mean is not easily understood and is difficult to compute.

ii) It gives greater importance to small items.

iii) It cannot be calculated when any observation is zero.

Relation among Arithmetic Mean (A), Geometric Mean (G) and Harmonic Mean (H):

i) $G = \sqrt{A \times H}$

ii) $A \geq G \geq H$, equality holds when all observations have same magnitude.

Median: The median of distribution is that value which divides it into two equal parts. Median is a positional average and is the suitable measure of central tendency when individuals are ranked on the basis of some qualitative characteristics such as intelligence, poverty etc., which cannot be measured numerically. It is also used when

open-ended classes are given at one or both the ends or data arise from some skewed distribution such as income distribution.

Median is computed by arranging the observations in ascending or descending order of magnitude. If n, the number of observations is odd, then the size of $(n+1)/2^{th}$ observation will be the median and if n is even then average of two middle observations is the median.

Median for discrete frequency distribution is the observation corresponding to the cumulative frequency (less than type) just greater than N/2.

For a grouped frequency distribution the class corresponding to the cumulative frequency just greater than N/2 is called the median class and median is determined by the formula:

$$M_d = L + \frac{h}{f}(N/2 - C)$$

where L: is the lower limit of the median class

f: is the frequency of the median class

h: is the size of the median class

C: is the cumulative frequency of the class preceding the median class

N $= \Sigma f$

Merits: It is rigidly defined, easy to calculate and easy to understand. It is a positional average and hence not affected by extreme values. It can be calculated for open-ended distribution.

Demerits: Median is not based on all the observations. It is not suitable for further treatment and is more sensitive to the fluctuations of sampling.

Mode: Mode is defined as the value, which occurs most frequently in a set of observations and around which other observations of the set cluster densely. Mode of a distribution is not unique. If two observations have maximum frequency then the distribution is bimodal. A distribution is called multi-modal if there are several values that occurs maximum number of times. Mode is often used where we need the most typical value e.g. in business forecasting the average required by the manufacturers of the sizes of readymade garments, shoes etc.

Mode of a discrete frequency distribution can be located by inspection as the variable value corresponding to maximum frequency. For a continuous frequency distribution we first calculate the modal class and then mode is determined by the following formula:

$$\text{Mode} = L + h \times \frac{f_m - f_1}{2f_m - f_1 - f_2}$$

Here L, h, f_m, f_1 and f_2 are respectively the lower limit of the modal class, the size of modal class, frequency of the modal class, frequency preceding the modal class and frequency following the modal class.

If there are irregularities in the distribution or the maximum frequency is repeated or the maximum frequency occurs in the very beginning or at the end of the distribution then the mode is determined by grouping method.

Merits: Mode is not affected by extreme observations and can be calculated for open ended distributions.

Demerits: It is ill defined, not based on all the observations, not unique and often does not exit.

Relationships among Mean, Mode and Median:

i) For a symmetrical distribution Mean = Median = Mode

ii) For a skewed distribution Mean - Mode = 3(Mean ï Median) or
Mode = 3 Median ï 2 Mean

iii) For a positively skewed distribution Mean > Median >Mode

iv) For a negatively skewed distribution Mean < Median <Mode

Partition Values or Quantiles: Some times we not only need the mean or middle value but also need values which divide the data in four or ten or hundred equal parts.

Quartiles: Are the three values which divide data into four equal parts. These are denoted by Q_1, Q_2, Q_3

On the lines of median formula, quartiles are given by

$$Q_i = L + \frac{iN/4 - C}{f} \times h \;\; ; \;\; i = 1, 2, 3$$

where L, C and f are lower limit, C.F. of preceding class and frequency of the class containing that quartile. Q_1 and Q_3 are called lower and upper quartiles respectively and Q_2 is equal to the median.

Deciles: Are the nine values which divide the data into 10 equal parts and deciles are denoted by D_1, D_2, ·····, D_9; where

$$D_i = L + \frac{iN/10 - C}{f} xh \;\; ; \;\; i = 1, 2, ..., 9$$

Percentiles: Are the ninety nine values which divide the data into 100 equal parts and are denoted by P_1, P_2, ·····, P_{99}; where

$$P_i = L + \frac{iN/100 - C}{f} xh \;\; ; \;\; i = 1, 2, ..., 99$$

Percentiles help to find the cut off values when the data be divided into a number of categories and per cent of observations in various categories are given.

1.5 Measures of Dispersion:

For comparing two sets of data or distributions, comparison of average values may not give complete picture as it may be possible that two sets of data or distributions may have the same mean but they may differ in dispersion or scatter or spread. Hence we must compare scatter/dispersion in addition to comparison of locations or means.

The degree to which numerical data tend to scatter or spread around the central value is called dispersion or variation. Further, any quantity that measures the degree of spread or scatter around the central value is called measure of dispersion. The various measures of dispersion are:

i) Range

ii) Quartile Deviation

iii) Mean Deviation

iv) Variance

v) Standard Deviation

vi) Quartile Coefficient of Dispersion

vii) Coefficient of Mean Deviation

viii) Coefficient of Variation

Absolute and Relative Measures of Dispersion:

A measure of dispersion indicates the degree to which the numerical data tend to spread or scatter around an average value. The measures expressed in terms of the units of the data are called absolute measures. Range, quartile deviation, mean deviation and

standard deviation are the common examples of absolute measures. The measures, which are independent of the units of measurement are called the relative measures. These are pure numbers and often expressed as percentages. Quartile coefficient of dispersion, coefficient of mean deviation and coefficient of variation are a few examples of relative measures.

Range: It is defined as the difference of the two extreme observations of a data set. Range is used when we need a rough comparison of two or more sets of data or when the observations are too scattered to justify the computation of a more precise measure of dispersion.

Merits and Demerits of Range: Range is rigidly defined and is the simplest measure of dispersion. It is also easy to interpret and calculate. Range is a crude and unreliable measure of dispersion and it being based only on two extreme observations and has greater chances of being affected by fluctuations in sampling.

Quartile Deviation or Semi-Quartile Range (QD): It is mathematically defined as

$$QD = (Q_3 - Q_1)/2$$

Quartile deviation is preferred when distribution is skewed or open-ended. It is also used when error caused by extreme values is to be minimized.

Merits and Demerits of QD: It is rigidly defined and easy to calculate and understand. It is not affected by the extreme values. The main demerits of quartile deviation are that it is neither based on all the observations nor suitable for further mathematical treatment. It is also sensitive to the fluctuations in sampling.

Mean Deviation (MD): The arithmetic mean of the absolute deviations about any point A is called the mean deviation or mean absolute deviation about the point A. The point óAô may be taken as mean, median or mode of the distribution. It is a useful measure of dispersion in business and economics when extreme observations influence the standard deviation unduly.

The MD of n observations $x_1, x_2, x_3, \cdots\cdots, x_n$ about any point A is given by

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$$

If a frequency distribution is given then,

$$MD = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - A| \qquad \text{where } N = \sum f_i$$

Merits: Mean deviation is rigidly defined, based on all the observations, easy to calculate and relatively easy to interpret. It is not affected much by the extreme values.

Demerits: Mean deviation is not suitable for further mathematical treatment. It also ignores the signs of deviations and hence creates some artificiality in the result.

Variance: The arithmetic mean of the squared deviations taken about mean of a series is called variance ($\mathring{u}^2$). Mathematically, variance is defined as:

$$\mathring{u}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^{\,2} - \bar{x}^2$$

For a frequency distribution, the variance is given by:

$$\mathring{u}^2 = \frac{1}{N}\sum f_i(x_i - \bar{x})^2 = \frac{1}{N}\sum f_i x_i^{\,2} - \bar{x}^2 \qquad \text{where } N = \sum f_i \text{ and } \bar{x} = \frac{\sum f_i x_i}{N}$$

Standard Deviation: The positive square root of the arithmetic mean of the squared deviations of observations in a data series about its arithmetic mean is called standard deviation ($\mathring{u}$) i.e. it is the square root of variance i.e.

$$\mathring{u} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum x_i^{\,2} - \bar{x}^2} \text{ and for a frequency distribution}$$

$$\mathring{u} = \sqrt{\frac{1}{N}\sum f_i(x_i - \bar{x})^2} = \sqrt{\frac{1}{N}\sum f_i x_i^{\,2} - \bar{x}^2}$$

Standard deviation is a stable measure and is regarded as the best and most powerful measure of dispersion.

Merits and Demerits of Standard Deviation:

Standard deviation is rigidly defined and based on all the observations. It is relatively less sensitive to the sampling fluctuations and also suitable for further mathematical treatment. Standard deviation is independent of change of origin but not of scale. The main demerit of standard deviation is that it gives greater weightage to extreme values and cannot be calculated in case of open-ended classes.

Shortcut method for computing AM and variance

AM and Variance are the most frequently used measures of central tendency and dispersion respectively and they are also used in finding the C.V. which is a relative measure of dispersion. Shortcut method is given in the following steps:

i) Find the mid values of class intervals if not given. Let $x_1, x_2, \cdots, x_n$ be the mid values.

ii) For every class interval find $u_i = (x_i$ ï $A)/h$ where A = assumed mean and h = width of the class interval.

iii) Find $\Sigma f_i u_i$ and $\Sigma f_i u_i^2$ and

$$\text{AM}(\bar{x}) = A + h\,\bar{u} = A + \frac{\sum f_i u_i}{N} \text{ x } h$$

$$\text{Variance } (\sigma_x^2) = h^2\left[\sum f_i u_i^2 / N - (\sum f_i u_i / N)^2\right]$$

Relative Measures of Dispersion:-

i) Quartile Coefficient of Dispersion (QCD) = $\frac{Q_3 - Q_1}{Q_3 + Q_1} x100$

ii) Coefficient of Mean Deviation = $\frac{\text{MD about A}}{A} x100$, where A = mean or median or mode

iii) Coefficient of Variation (CV) = $\frac{SD}{\text{Mean}} x\,100$

Relative measures of dispersion are used while comparing the variability of series having different or same units of measurement. They can also be used for comparing two or more series for consistency. A series with smaller coefficient of dispersion is said to be less dispersed or more consistent (or homogeneous).

Combined Standard Deviation: Combined standard deviation of two groups is denoted by σ_p and is computed follows:

$$\mathring{u}_p = \sqrt{\frac{n_1\mathring{u}_1^2 + n_2\mathring{u}_2^2 + n_1 d_1^2 + n_1 d_2^2}{n_1 + n_2}}$$

where σ_1 = SD of first group

σ_2 = SD of second group

$d_1 = (\overline{X}_1 - \overline{X}_p)$; $d_2 = (\overline{X}_2 - \overline{X}_p)$

$\overline{X}_1$, $\overline{X}_2$, $\overline{X}_p$ are the group means and pooled mean respectively.

The above formula can be extended to find the combined SD of three or more groups.

1.6 Skewness and Kurtosis:

Skewness: Literally means lack of symmetry, skewness gives an idea about the shape of the curve that can be drawn with the help of the given data. The frequency curve of a skewed distribution is not symmetrical but stretched more to one side than to the other.

Skewness is often measured by the Karl-Pearson Coefficient of skewness defined as:

$$S_k = \frac{3(\text{Mean - Median})}{\text{Standard Deviation}} = \frac{\text{Mean - Mode}}{\text{Standard Deviation}}$$

Since (M ï M_d)/ů lies between ± 1, hence S_k lies between ± 3.

While in terms of moments $ƃ_1 = \frac{ε_3^2}{ε_2^3}$ and $ɔ_1 = \sqrt{ƃ_1}$

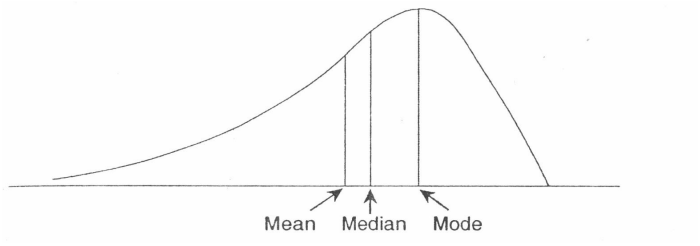
where $\mu_r = (1/N)\sum f_i (x_i - \bar{x})^r$ is r^{th} order central moment of the variable X and sign of $ɔ_1$ is same that of μ_3.

Graphically it can be shown as under:

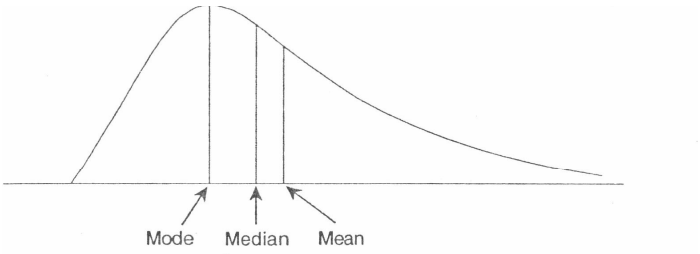
If $ɔ_1 = 0 \Rightarrow$ curve is normal (symmetrical)

$ɔ_1 < 0 \Rightarrow$ curve is skewed to the left (negative skewness)

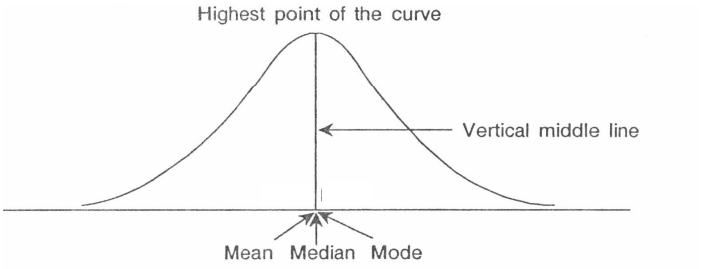
$ɔ_1 > 0 \Rightarrow$ curve is skewed to the right (positive skewness)



Graph showing negative skewness (the curve stretched more to the left)



Graph showing positive skewness (the curve stretched more to the right)



Graph showing normal (symmetrical) curve

Remarks:

i) For a symmetrical distribution mean, mode and median are equal

ii) Skewness is a pure number having no units of measurement and thus can be used to compare the skewness in sets of data with same or different units of measurements.

iii) For a positively skewed distribution AM > Median > Mode and for a negatively skewed distribution AM < Median < Mode.

Kurtosis: The flatness or peakedness of top of the curve is called kurtosis, which is also an important characteristic of a distribution. Kurtosis is measured in terms of moments and is given by:

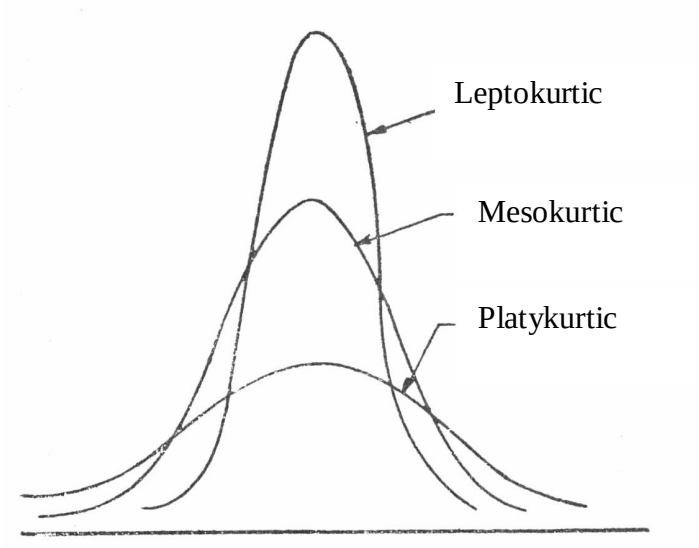
$\beta_2 = \mu_4 / \mu_2^2$ and $\gamma_2 = \beta_2 - 3$

For normal distribution curve $\beta_2 = 3$ (or $\gamma_2 = 0$) and flatness or peakedness of the curve of a distribution is compared in relation to normal curve and the distributions have been classified as follows:

If $\beta_2 = 3 \Rightarrow \gamma_2 = 0$, then the curve is properly peaked, i.e. **Mesokurtic** curve.

$\beta_2 > 3 \Rightarrow \gamma_2 > 0$, then the curve is more peaked than normal, i.e. **Leptokurtic** curve

$\beta_2 < 3 \Rightarrow \gamma_2 < 0$, then the curve is less peaked than normal, i.e. **Platykurtic** curve



Graph showing Mesokurtic, Leptokurtic and Platykurtic Curves

Coefficient of Kurtosis is also a pure number having no units of measurements and thus can be used to compare the flatness of the top of curves for frequency distributions with same or different units.

Example-4: Calculate the AM, median, Q_1, D_4, P_{55}, mode, G.M. and H.M. for the data on salaries of 1000 employees in a company.

Salaries (000Rs)	**No. of employees (f)**	**Class mark (x)**	**fx**	**c.f.**	**log x**	**f log x**	**f/x**
1-3	50	2	100	50	0.3010	15.50	25
3-5	110	4	440	160	0.6021	66.23	27.5
5-7	162	6	972	322	0.7782	126.07	27.00
7-9	200	8	1600	522	0.9031	180.62	25.00
9-11	183	10	1830	705	1.0000	183.00	18.3
11-13	145	12	1740	850	1.0792	156.48	12.1
13-15	125	14	1750	975	1.1461	143.33	8.9
15-17	15	16	240	990	1.2041	18.06	0.9
17-19	8	18	144	998	1.2553	10.04	0.4
19-21	2	20	40	1000	1.3010	2.60	0.1
Total	**N = 1000**		**8856**			**901.93**	**145.2**

Solution:

(i) AM ($\overline{X}$) = $\sum fx/N = 8856/1000 = 8.856$ = Rs. 8856

(ii) Median = $L + \frac{N/2 - C}{f} \times h = 7 + \frac{500 - 322}{200} \times 2 = 8.78$ = Rs.8780

(iii) Lower Quartile (Q_1) = $L + \frac{N/4 - C}{f} \times h = 5 + \frac{250 - 160}{162} \times 2 = 6.111$ = Rs.6111

(iv) 4th decile (D_4) = $L + \frac{4N/10 - C}{f} \times h = 7 + \frac{400 - 322}{200} \times 2 = 7.78$ = Rs.7780

(v) 55^{th} percentile (P_{55}) = $L + \frac{55N/100 - C}{f} \times h = 9 + \frac{550 - 522}{183} \times 2 = 9.306$ = Rs. 9306

(vi) Mode = $L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h = 7 + \frac{200 - 162}{2(200) - 162 - 183} \times 2 = 8.382$ = Rs.8382

(vii) GM = Antilog ($\sum f \log x/N$) = Antilog (901.93/1000)

= Antilog (0.901) = 7.979 = Rs. 7979

(viii) HM = Reciprocal of ($1/N \sum f/x$) = Rec. of (145.2/1000) =1000/145.2

= 6.887 = Rs. 6887

Example-5: Find the average rate of increase in population which in the first decade has increased by 20%, in the second decade by 30% and in the third decade by 40%. Find the average rate of increase in the population.

Solution: To find the average rate of growth, geometric mean is the appropriate average.

Decade	% Rise	Population at the end of the decade (x)	log x
1^{st}	20	120	2.0792
2^{nd}	30	130	2.1139
3^{rd}	40	140	2.1461
			Ælog x = 6.3392

GM = Antilog (1/n Ælog x) = Antilog (6.3392/3) = Antilog (2.1130) = 129.7

Thus average rate of increase in the population is (129.7 ï 100) = 29.7 per cent per decade.

Example-6: A taxi driver travels from plain to hill station 100 km distance at an average speed of 20 km per hour. He then makes the return trip at average speed of 30 km per hour. What is his average speed over the entire distance (200 km)?

Solution: To find the average speed, harmonic mean is an appropriate average. Harmonic mean of 20 and 30 is:

$$\text{Harmonic mean} = \frac{2}{\frac{1}{20}+\frac{1}{30}} = \frac{2}{\frac{5}{60}} = \frac{2 \text{ x } 60}{5} = 24 \text{ km per hour}$$

Example-7: Find the mean deviation about AM and hence find the coefficient of mean deviation.

Daily wages (Rs.)	0-20	20-40	40-60	60-80	80-100	Total
No. of wage earners (f)	2	5	10	10	5	**N=32**
Mid value (x)	10	30	50	70	90	
fx	20	150	500	700	450	**1820**
$\lvert x - \bar{x} \rvert$	46.9	26.9	6.9	13.1	33.1	
$f\lvert x - \bar{x} \rvert$	93.8	134.5	69	131	165.5	**593.8**

AM ($\bar{x}$) = $\sum fx/N$ = 1820/32 = 56.9

MD = $\sum f\lvert x - \bar{x} \rvert /N$ = 593.8./32 = 18.55

Coefficient of $\text{MD} = \frac{\text{MD}}{\text{AM}} \times 100 = \frac{18.55}{56.9} \times 100 = 32.6\%$

Example-8: The runs scored by two batsmen X and Y in 10 innings are given below. Find out which is better runner and who is more consistent player?

											Total
X	90	110	5	10	125	15	35	16	134	10	**550**
Y	65	68	52	47	63	25	25	60	55	60	**520**
$x - \bar{x}$	35	55	-50	-45	70	-40	-20	-39	79	-45	
$(x - \bar{x})^2$	1225	3025	2500	2025	4900	1600	400	1521	6241	2025	**25462**
$y - \bar{y}$	13	16	0	-5	11	-27	-27	8	3	8	
$(y - \bar{y})^2$	169	256	0	25	121	729	729	64	9	64	**2166**

Series X: $\bar{X} = \sum x/n = 550/10 = 55$

$$\mathring{u}_x^2 = \sqrt{\Sigma(x-\bar{x})^2/n} = \sqrt{25462/10} = \sqrt{2546.2} = 50.46$$

$$CV = \frac{\mathring{u}_x}{\bar{X}} \times 100 = \frac{50.46}{55} \times 100 = 91.74\%$$

Series Y: $\bar{Y} = \sum y/n = 520/10 = 52$

$$\mathring{u}_y^2 = \sqrt{\Sigma(y-\bar{y})^2/n} = \sqrt{2166/10} = \sqrt{216.6} = 14.71$$

$$CV = \frac{\mathring{u}_y}{\bar{Y}} \times 100 = \frac{14.71}{52} \times 100 = 28.28\%$$

Conclusion: X is better runner since $\bar{X}$ = 55 is more than $\bar{Y}$ = 52 and Y is more consistent player since CV in Y series is less than CV in X series.

Example 9: The profits (in million rupees) earned by 70 companies during 2008-09 are given below. Compute the AM, SD and CV

Profit	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**Total**
No. of Companies (f)	6	20	24	15	5	**70**
Class mark (x)	15	25	35	45	55	
$(x - \bar{x})$	-19	-9	1	11	21	
$(x - \bar{x})^2$	361	81	1	121	441	
$f(x - \bar{x})^2$	2166	1620	24	1815	2205	**7830**
fx	90	500	840	675	275	**2380**
fx^2	1350	12500	29400	30375	15125	**88750**
u=(x-35)/10	-2	-1	0	1	2	
fu	-12	-20	0	15	10	**-7**
fu^2	24	20	0	15	20	**79**

AM ($\bar{x}$) = $\sum fx/N$ = 2380/70 = 34 million rupees

Method-I (when AM is a fraction)

$$\text{Variance } (\sigma_x^2) = \frac{1}{N}\Sigma fx^2 - \left(\frac{1}{N}\Sigma fx\right)^2 = 88750/70 - 34^2 = 111.86$$

$$\text{SD } (\sigma_x) = \sqrt{\text{Variance}} = \sqrt{111.86} = 10.58 \text{ million rupees}$$

Method-II (when AM is an integer)

$$\text{Variance } (\sigma_x^2) = \frac{1}{N}\Sigma f(x - \bar{x})^2 = 7830/70 = 111.86$$

$$\text{SD } (\sigma_x) = 10.58$$

Method –III (Shortcut method)

Consider u = (x-A)/h where A (assumed mean) = 35 and h (width of class interval) = 10

$$AM = A + \frac{\Sigma fu}{N} \times h = 35 + \left(\frac{-7}{70}\right) \times 70 = 35 - 1 = 34$$

Variance (σ_x^2) = $h^2 [\Sigma fu^2/N - (\Sigma fu/N)^2] = 100[79/70 - (-7/70)^2]$

= $100 [1.1286 - 0.01] = 100(1.1186) = 111.86$

SD $(\sigma_x) = \sqrt{111.86}$ = 10.58

Example-10: The data regarding daily income of families in two villages are given below:

	Village A	**Village B**
Number of families	600	500
Average income (Rs.)	175	186
Variance (Rs^2)	100	81

i) In which village there is more variation in income.

ii) What is the combined standard deviation of income of two villages?

Solution:

i) Village A CV = $\frac{SD}{AM} \times 100 = \frac{10}{175} \times 100 = 5.71\%$

Village B CV = $\frac{9}{186} \times 100 = 4.84\%$

Since CV in village A is more than village B, therefore, there is more variation of income in village A.

ii) The Combined SD of income in both the villages A and B is given by

$$\mathring{u}_p = \sqrt{\frac{n_1\mathring{u}_1^2 + n_2\mathring{u}_2^2 + n_1d_1^2 + n_2\mathring{u}_2^2}{n_1 + n_2}}$$

Here $d_1 = \overline{X}_1 - \overline{X}_p$; $d_2 = \overline{X}_2 - \overline{X}_p$

$$\overline{X}_p \text{ (pooled mean)} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2} = \frac{600x175 + 500x186}{600 + 500}$$

$$= \frac{105000 + 93000}{1100} = \text{Rs.}180$$

$d_1 = |175 \text{ï } 180| = 5$; $d_2 = |186 \text{ï } 180| = 6$

$$\mathring{u}_p = \sqrt{\frac{600x100 + 500x81 + 600(25) + 500(36)}{1100}} = \sqrt{\frac{133500}{1100}}$$

$$= \sqrt{121.36} = \text{Rs.}11.02$$

1.7 Diagrammatic and Graphical Representation of Data:

Numerical data may be represented in a simple and attractive manner in the form of diagrams and graphs. Diagrammatic representation is used when data relating to different times and places are given and are independent of one another. Graphical representation is used when we have to represent the data of frequency distribution or of a time series. Diagrammatic representation includes bar diagram, rectangular diagram and pie diagram, whereas a graphical representation includes frequency graphs such as histogram, frequency polygon, and ogive.

Simple Bar Diagram:

It is diagram of rectangles where each rectangle has some value and with following features:

i) Length or height of the bar varies with value of the variable under study.

ii) Bars are of the same width, equidistant from each other and are based on the same line.

Rectangular Diagram: In a rectangular diagram both length as well as width of the rectangles, are taken into consideration. The rectangular diagram is also called two-dimensional diagram. This diagram is used when two sets of data with different subdivisions are to be compared.

Pie Diagram: A pie diagram is circle divided into sectors indicating the percentages of various components, such that the areas of these sectors are proportional to the shares of components to be compared. Pie diagram is useful when percentage distribution is presented diagrammatically.

Histogram: A histogram is a two dimensional diagram used to represent a continuous frequency distribution. For drawing a histogram we first mark along the x-axis all the class intervals and frequencies along the y-axis according to a suitable scale. With class intervals as bases we draw rectangles whose areas are proportional to the frequencies of the class intervals. If the class intervals are equal then the length of rectangles will be proportional to the corresponding class frequencies.

Frequency Polygon and Frequency Curve: A frequency polygon is obtained if we plot mid values of the class intervals against frequencies and the points are joined by means of straight lines. If we join one mid value before the first group and one mid value after the last group then areas under the frequency polygon and the corresponding histogram are equal. A frequency polygon can also be obtained by joining the mid points of the upper sides of rectangles in the histogram along with one mid value before the first class interval and one mid value after the last class interval.

A frequency curve is obtained if we join the mid points by means of a free hand curve. Frequency polygon is observations used to represent a discrete frequency distribution.

Ogive: An ogive is a cumulative frequency curve in which cumulative frequencies are plotted against class limits. There are two types of ogives.

Less Than Ogive: First we form a less than type cumulative frequency distribution. We then plot the upper limits of the classes along x-axis and the less than cumulative frequencies along y-axis. These points are joined by a free hand curve. This curve is called less than ogive or less than cumulative frequency curve.

More Than Ogive: First we form a more than type cumulative frequency distribution. Then we plot the lower limits of the classes along x-axis and more than cumulative frequency along the y-axis. The points are joined by free hand curve. This curve is known as ôMore than cumulative frequency curve or more than ogive.

Graphical Representation: Graphs are used to represent a frequency distribution which makes the data understandable and attractive. It also facilitates the comparison of two or more frequency distributions. In addition to it, some of the statistical measures like mode, median and partition values can be located through graphs. The following types of graphs are generally used.

Limitations of Graphical Representation: The graphs cannot show all those facts, which are available in the tables. They also take more time to be drawn than the tables.

Graphical Location of Mode:

Mode in a frequency distribution is located graphically by drawing the histogram of the data. The steps are:

i) Draw a histogram of the given data

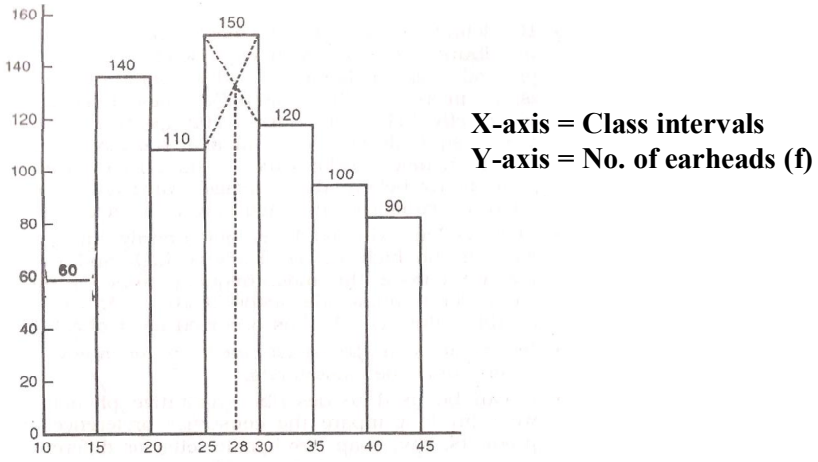
ii) Draw two lines diagonally in the inside of the modal class starting from each upper corner of the bar to the upper corner of the adjacent bar

iii) Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis which gives the modal value.

Example-11: Draw a histogram for the following distribution of earhead weights and locate the modal weight of the earheads from histogram and check by direct calculation.

Weight (in gms)	**10-15**	**15-20**	**20-25**	**25-30**	**30-35**	**35-40**	**40-45**
No. of earheads (f)	60	140	110	150	130	100	90

Solution: The histogram of this data is given below



Thus modal weight of earheads = 28 gms

Direct Calculation: Mode lies in the class 25-30

Modal Weight $= L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h = 25 + \frac{150 - 110}{2(150) - 110 - 120} \times 5$

$$= 25 + \frac{40}{70} \times 5 = 25 + 2.86 = 27.86 \simeq 28 \text{ gms}$$

Graphical Location of Median and other Partition Values: Median is graphically located from the cumulative frequency curve i.e. ogive. The various steps are:

Step-I: Draw less than or more than cumulative frequency curve i.e. Ogive.

Step-II: Mark a point corresponding to N/2 on the frequency axis (i.e. y-axis) and from this point draw a line parallel to x- axis which meets the Ogive at the point A (say).

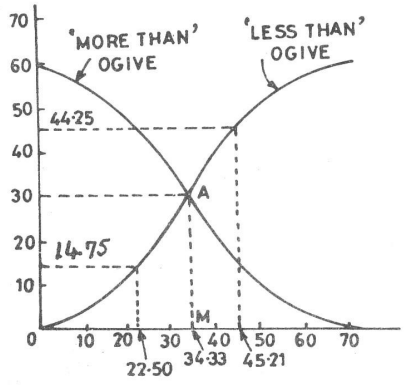
Step-III: From A, draw a line parallel to y-axis, which meets the x-axis at the point M (say). Then distance of the point M from the origin on the x- axis is the median.

Note: The other partition values viz. quartiles, deciles and percentiles can also be located graphically as described above.

Example-12: The bonus (%) announced by 59 steel industries in India is given below. Draw the less than and more than Ogives and locate the median, Q_1 and Q_3.

Bonus (%)	No. of industries	Less than c.f.	More than c.f.
0-10	4	4	59
10-20	8	12	55
20-30	11	23	47
30-40	15	38	36
40-50	12	50	21
50-60	6	56	9
60-70	3	59	3

Solution: The less than and more than Ogives are drawn below



X-axis: Bonus (%)
Y-axis: No. of industries

To locate the median, mark a point corresponding to N/2 along the Y-axis. At this point, draw a line parallel to X-axis meeting the Ogive at the point A. From A draw perpendicular to the X-axis meeting at M. The abscissa of M gives the median.

For quartiles Q_1 and Q_3, mark the points along the Y-axis corresponding to N/4 and 3N/4 and proceed as above. From the graph we see that

Median = 34.33, Q_1 = 22.50 and Q_3 = 45.21. Other partition values viz. deciles and percentiles can be similarly located.

Remark: Median can also be located by drawing a perpendicular from the point of intersection of the two Ogives as shown.

1.8 Box Plot (or Box-Whisker Diagram):

Box plot introduced by õTukeyô is a graphical representation of numerical data and based upon the following five number summary:

i) Smallest observation (Sample minimum)

ii) Lower quartile (Q_1)

iii) Median (M or Q_2)

iv) Upper quartile (Q_3)

v) Largest observation (Sample maximum)

It is an excellent tool for conveying valuable information about some descriptive measures like central tendency, dispersion, skewness etc. in data sets.

Box plot has a scale in one direction only. A rectangular box is drawn extending from the lower quartile (lies on the lower side of the rectangle) to the upper quartile (lies on the upper side of the rectangle). Thus the box plot represents the middle 50% of the data. The horizontal line within the box represents the median value. The vertical lines (whiskers) are then drawn extending from above and below the box to the largest and smallest values, respectively. Thus, the relative positions of the components illustrate how the data is distributed.

Box plots display differences between populations without making any assumption of the underlying statistical distribution and hence they are non-parametric. The spacing between the different parts of the box indicates the degree of dispersion and skewness present in data and identify outliers.

Single and Multiple Box Plots: A single box lot can be drawn for one data set. Alternately multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set. For a single box plot, the length of the box is arbitrary. For multiple box plots, the length of the box can be set proportional to the number of observations in a given group or sample. The stepwise procedure is given below:

i) Take the response variable on the vertical axis and factor of interest on the horizontal axis.

ii) Calculate M, Q_1, Q_3 and locate the largest and smallest observations.

iii) Draw the rectangular box with Q_3 and Q_1 lies on the upper and lower side of the box respectively.

iv) Draw the vertical lines (whiskers) extending from the upper and lower sides of the box to the largest and smallest values.

Detection of Outliers and Skewness: In refined box plots, the whiskers have a length not exceeding 1.5 x inter-quartile length i.e. 1.5 x (Q_3 ï Q_1). Any values beyond the ends of the whiskers are detected as outlier.

Skewed to the Left: If the box plot shows the outliers at the lower range of the data (below the box), the mean (+) value is below the median, the median line does not evenly divide the box and the lower tail of box plot is longer than the upper tail, than the distribution of data may be skewed to the left or negatively skewed.

Skewed to the Right: If the box plot shows the outliers at the upper range of the data (above the box), the mean (+) value is above the median, the median line does not evenly divide the box, and the upper tail of the box plot is longer than the lower tail, then the distribution of data may be skewed to the right or positively skewed.

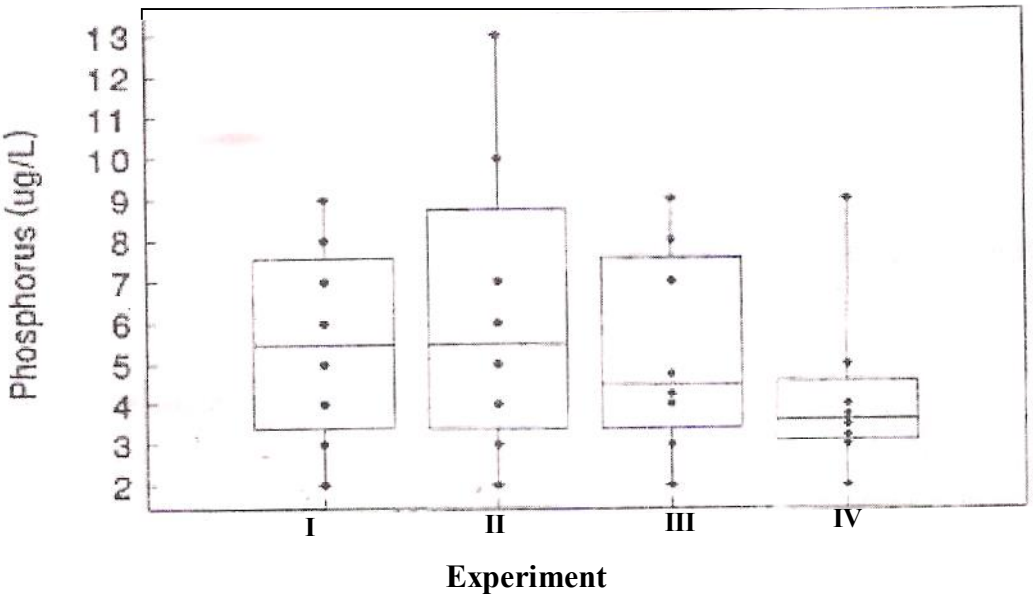
Example-13: Consider the following data on phosphorus (μg/l) recorded in four experiments:

Experiment	**Observations**
I	2, 3, 4, 5, 6, 7, 8 and 9
II	2, 3, 4, 5, 6, 7, 10 and 13
III	2, 3, 4, 4.25, 4.75, 7, 8 and 9
IV	2, 3, 3.25, 3.5, 3.75, 4, 5 and 9

Draw the box plots of the above data sets describing the relative positions of median, quartiles, maximum and minimum observations and draw your conclusions about the distribution of data.

Solution:

Box plots of four data sets (I-IV)



Experiment-1: Here the median = 5.5, Maximum and minimum values are 2 and 9 giving a range of 7. It is a symmetrical distribution since

i) The maximum and minimum observations are at equal distances from the median

ii) The maximum and minimum observations are at equal distances from the box

iii) Median lies exactly in the centre of the box

Experiment-II: Here the median is = 5.5, Maximum and minimum values are 2 and 13 giving a range of 11. It is a skewed distribution since

i) The maximum and minimum observations are not at equal distances from the median

ii) The maximum and minimum observations are not at equal distances from the box

iii) Median does not lie exactly in the centre of the box

Experiment-III: Here the median is = 4.5, Maximum and minimum values are 2 and 9 giving a range of 9. It is a skewed distribution since

i) The maximum and minimum observations are not at equal distances from the median

ii) The maximum and minimum observations are not at equal distances from the box

iii) Median does not lie exactly in the centre of the box

iv) The values 4, 4.25 and 4.75 are being clumped together.

Experiment-IV: Here the median is = 3.375, Maximum and minimum values are 2 and 9 giving a range of 7. It is a skewed distribution since

i) The maximum and minimum observations are not at equal distances from the median

ii) The maximum and minimum observations are not at equal distances from the box

iii) Median does not lie exactly in the centre of the box

iv) Clumping of low values (as in III case)

EXERCISES

1. The mean marks in statistics of 100 students of a class was 72. The mean marks of boys was 75, while their number was 70. Find the mean marks of girls in the class.

2. The following data give the electricity consumed (K.watt) by 100 families of Hisar

Electricity consumption	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**
No. of users	6	25	36	20	13

Calculate AM, median, mode, S.D., C.V. and find the range with in which middle 50% consumers fall [Hint: Range of middle 50% = Q_3 ï Q_1]

3. The mean and standard deviation of a set of 100 observations were worked out as 40 and 5 respectively by a computer which by mistake took the value 50 in place of 40 for one observation. Find the correct mean and variance.

4. Calculate S.D. and C.V. from the following data

Profits (10^7Rs.)	**Less than 10**	**Less than 20**	**Less than 30**	**Less than 40**	**Less than 50**	**Less than 60**
No. of companies	8	20	40	70	90	100

[Hint: First convert the cumulative frequencies in to simple frequencies]

5. For the following distribution of heights of 80 students in a class, find AM and S.D. by shortcut method and hence find C.V.

Heights (cm)	**150-155**	**155-160**	**160-165**	**165-170**	**170-175**	**175-180**	**180-185**
No. of students	6	9	20	23	15	5	2

6. Draw histogram, frequency polygon, frequency curve, less than and more than ogives and box plot of the following frequency distribution. Locate the mode, median, Q_1 and Q_3 from the appropriate graph.

Marks	40-50	50-60	60-70	70-80	80-90
No. of students	10	20	40	15	5