Module on Business Statistics

For Centurion University

Date: 23/01/2018

UNIT I **COLLECTION, CLASSIFICATION AND PRESENTATION**

OF STATISTICAL DATA AND THEORY OF SAMPLING

1.1 Objectives: After going through this chapter we shall be able to know what the term statistics means and how different data are collected and they are represented.

1.2 Introduction: Statistics can be defined in two ways. Statistics, in plural sense defined as set of facts and figures collected by an individual or an authority on the concerned topic. It comprises statistics or data on different topics like population, sports, birth-death etc. In singular sense, statistics is defined as the scientific methods which are meant for the collection, analysis and interpretation of numerical data like, a book of statistics gives different methods and measures to analyze the data.

1.3. Scope of Statistics: Statistics is extensively in the field Economics, Politics, Businesses and Management. We shall now discuss the scope of Statistics in these fields.

a. **Statistics and State:** The government of any country uses the data of the economy on different phenomena for its growth and development. For example, the data on crimes helps government to implement appropriate measures for better controlling of crime in the country. Nowadays a government is mainly interested in the well-being of its citizen, thus, it uses other statistics on several indicators, such as general price level, production, consumption and distribution of income among its people.

b. **Statistics in Economics:** Statistical analysis is widely used in the field of Economics. It provides solution to a variety of economic problems such as production, consumption and distribution. For example, an analysis of data on consumption may reveal the pattern of consumption of various commodities by different sections of the society. Similarly, data on prices, wages, profits, investments, and savings are very important in formulating economic policies.

c. **Statistics in Business and Management:** For a progressive business concern, analysis of costs, revenue, profits, labour and capital, marketing, etc. are essential. Business planning involves business forecasting based on market surveys on demand, availability of substitute goods, opinions of consumers regarding different brands, consumers' preferences, etc. Statistical methods are useful to business in formulating its business policies and activities in the field of production, finance, personnel, accounting and quality control. Modern businesses make extensive use of graphs, charts, and diagrams in their sale promotion efforts and display of their production achievements.

1.4. Limitations of Statistics: In the following section we shall discuss the limitations of statistics.

a. **Statistics is applicable only to quantitative data: Statistics** usually deals with data which are expressed in numerical terms. For qualitative data, we cannot use statistics directly. To deal with qualitative data, some numerical value has to be assigned to them in order to operate on them.

b. **It does not deal with individuals:** Statistics is a method use to study only the aggregates. A single or isolated figure cannot be regarded as statistics.

c. **Statistical Laws are not exact:** The statistical laws are based on certain conditions always some chance factor is associated with them for being true. These laws cannot be universally accepted like the laws of Physics and Chemistry.

d. **Statistical decisions are applicable only on the average:** The methods of statistics used to reveal only the average behavior of a phenomenon. It does not concern itself with individual variable.

e. **Statistics can be misused:** Statistical methods, if not applied in the proper perspective of the collected data, may lead to false conclusions. These should therefore be handled with utmost care and by experts only.

In the following section we shall discuss different types of data.

1.5. Types of data and its collection: Data are systematic record of values taken by a variable or a number of variables on a particular point of time or over different points of time. Statistical data are of two types viz. 'primary' and 'secondary'. Primary data are those data which are collected directly from the field foe some specific purpose and hence is original in nature. These data are published by authorities who themselves are responsible for their collection. On the other hand, data collected by someone but used by another or collected for one purpose and used for another are called secondary data. Following are few examples of primary and secondary data.

Primary data

1) "Reserve Bank of India Bulletin", published monthly by Reserve Bank of India.

2) "Indian Textile Bulletin", issued monthly by Textile Commissioner, Mumbai.

3) "Monthly Coal Bulletin", issused monthly by the Office of the Chief Inspector of Mines, Dhanbad.

Secondary data

1) "Monthly Abstract of Statistics" published by Central Statistical Organisation (C.S.O), Government of India, New Delhi.

2) “Statistical Abstract of the Indian Union”, issued by C.S.O, Mumbai

Data may be quantitative or qualitative in nature. Data of quantitative nature are technically called variables whereas data of qualitative nature are called attributes. For example, the number of student of Presidency University is 500, and then it is quantitative while religion of these students is qualitative. For more example we can consider, heights and weights of a group of student, agriculture production of rice in one year or over a period of time are quantitative variables whereas hard work, beauty, honesty etc. are attributes. Again variable can be discrete or continuous. Suppose we are considering number of students in different college, then these value are discrete in nature whereas the heights of these students can take any value within its range, hence this is continuous variable.

When data collected on a single point of time over different sections or on single section over different point of times, are known as cross-section data and when the data collected over the period of time is known as time series data.

1.6. Collection of Data: Data can be collected in two ways. One is *census method* and another is *sampling method*. Before going in detail about these methods let us first define population and sample. During the statistical investigation, the totality of statistical information on a particular character from all the members of a group is called population or universe; examples population of registered schools in a state, population of average income etc. The population may be finite or infinite.

The complete enumeration of the population is quite impractical because it consumes more money and time. Thus, a sample is taken out of population which can be used to analyze the characteristics of population. In sample method a part of population is considered and used to draw some conclusion on the characteristics of a population. When the data is collected on a population the method of collecting the data is known as census method and the method of collecting data on the basis of sample is known as sampling method. When the sample method is used the observer may arrive on some conclusions but these conclusions may inherent some error which is known as sampling error and these errors are unavoidable in any circumstances. But use of sampling method leads to considerable gains in term of cost and time and better handling of data.

1.7. Advantages of Sampling Method:

In this section we shall discuss some important advantages of sampling method.

i. **Lower cost and less time:** Since the sampling method considers only a portion of an entire population, it takes less staff and time and thus reduces both cost and time associated with it.
ii. **Better scope for information:** Since due to a time constraint in the census method, the surveyor cannot interact more closely with the household to get the information. Thus, using method of sampling, the surveyor can able to collect more information from the household.

iii. **Better quality of data:** In the census method, due to time constraint, we do not get good quality of data. But in sample survey, one can have a better quality of data as the survey consists of all the information related to objective of the study.

iv. **Detection of error:** For the population, we do not have a standard error, but for the sample we do have a standard error. Given the information of the sample mean and standard error, we can construct the limit within which almost all the sample value will lie.

v. The population in the census method or in complete enumeration can be infinite or hypothetical. The sample surveys to avoid these problems and is the best alternative to do a statistical analysis.

1.8. Some Concepts

i. **Parameter:** The statistical constant of the population such as Mean (μ), variance (σ^2), standard deviation (σ), moments (μ_r), skewness (β_1), kurtosis (β_2), correlation coefficient (R), etc., are known as ***parameters.*** It is important to note that the value of a parameter is computed from all the population observations. Thus, parameters are the function of the population values.

ii. **Statistic:** When the similar statistical constants computed for the sample drawn from the given population, such constants are known as ***statistics***. Statistics of a sample are denoted as Mean ($\bar{x}$), variance (s^2), standard deviation (s), moments (m_r), skewness (b_1), kurtosis (b_2), correlation coefficient (r), etc. Of course the statistics are the function of sample values as it is calculated on the basis of sample observations.

1.9. Random Sampling vs. Non Random Sampling

Sampling methods are majorly divided into two categories; random sampling and non-random sampling. In first case, each member of the sample has a fixed probability to be included in the sample whereas in non-random sampling there is no such chance of being included in the sample.

The important differences between these two methods of sampling are compiled in the next section.

i. Random sampling is a sampling technique, in which each member of the population gets an equal opportunity to be selected as representative sample. Whereas the non-random sampling is a sampling method, where it is not known which member of the population will be selected as a sample.

ii. The random sampling is also known as probability sampling as it is based on randomization or chance and since in non-random sampling probability or chance is not applied it is referred to as non-probability sampling.

iii. The probability of selection in random sampling is fixed and known while in non-random sampling the probability of selection is zero.

iv. The results derived from random sampling are free from bias while the results of non-random sampling are more or less biased.

v. Since the selection procedure of random sampling is based on probability, the extent to which it represents the whole population is higher as compared to the non-random sampling. That is why extrapolation of results to the entire population is possible in the random sampling but not in non-random sampling.

vi. To test a hypothesis, probability sampling is used while to generate a hypothesis non-probability sampling is used.

1.10. Types of Sampling: There are several ways of selecting a sample. We describe some of the important types of sampling.

i. Simple Random Sampling: Simple Random Sampling or simply Random Sampling is a method of selecting a group of units in such a manner that every member of the population has the equal chance of being included in the sample. There are two ways of drawing a simple random sampling. Firstly, if the unit selected in any draw is replaced before making the next draw, then the sampling method is called ***Simple Random Sampling with Replacement.*** Thus, in this case, the population remains the same before each drawing. Secondly, if the unit selected in any draw is not replaced before making the next draw, then the sampling method is known as ***Simple Random Sampling without Replacement.*** It is clear then; the size of population goes on diminishing as one by one unit gets selected.

ii. **Stratified Random Sampling:** When the population is heterogeneous with respect to the variable or characteristics under study, the whole population is divided into sub groups known as strata where units within each stratum are as homogeneous as possible. The division of population into several strata is known as stratification. When all the strata combined together, we get stratified sample and if the selection of strata is done by random sampling, the process is known as ***Stratified Random Sampling.*** If a proper method of stratified sampling is followed where each strata differs from another strata but the units within the strata is much homogenous, then it will provide a much better results than a random sampling.

iii. **Systematic Sampling:** In systematic sampling, the first sample unit is selected at random and the remaining units are selected on equal intervals after arranging the population in some order. If the population size is finite, the units can be arranged serially. From the first k of these, a single unit is chosen at random. This unit and every k-th unit thereafter constitute a ***Systematic Sample***. In order to obtain a systematic sampling of 50 villages out of 1000 in West Bengal, i.e., one out of 20 on an average, all the villages have to be numbered serially. From the first 20 of these a village is selected at random say serial number 15. Then the villages with serial numbers 15, 105, 195, 285, 375,... ... constitute the systematic sample.

iv. **Cluster Sampling:** In ***Cluster Sampling***, we divide the population into groups called clusters. Then we select a sample of clusters using a simple random sampling. The population units in each of the clusters are assumed to be as heterogeneous as those in the total population. That is, each cluster is itself a representative of the population.

The basic principles of cluster sampling are:

1. The differences or variability within a cluster should be as large as possible such that the variability within each cluster should be the same as that of the population.
2. The variability between the clusters should be as small as possible.

1.11. Sampling Distribution: A sampling distribution is the probability distribution of a given statistic based on a random sample. If for each sample of population, the value of the statistic is calculated, a series of values of the statistic will be obtained. If the number of sample is large, these may be arranged into frequency table. The frequency distribution of the statistic that would be obtained if the number of samples, each of the same size ('n'), were infinite is called the 'sampling distribution' of the statistic.

Like any other distribution, a sampling distribution may have its mean, standard deviations and moment of higher orders. Of particular importance is the standard deviation, which is designated as the 'standard error' of the statistic.

- There are two important sampling distribution
  - I. **Sampling Distribution of Sample Mean:** If $\bar{x}$ represents the mean of random sample of size *n*, drawn from a population of size *N* with mean μ and standard deviation σ then the sampling distribution of $\bar{x}$ is approximately a normal distribution with mean = μ and s.d = standard error of $\bar{x}$, provided that sample size *n* is sufficiently large.
  - II. **Sampling Distribution of Sample Proportion: If** *p* represents the proportion of smokers in a random sample of size *n* drawn from a lot with proportion of smokers *P*, then the sampling distribution of *p* is approximately a normal distribution with mean = *P* and s.d.= standard error of *p*, provided that sample size *n* is sufficiently large.
    - A sample size of 30 or more is regarded as large samples.

1.12. Standard Error (S.E): The standard deviation of the sampling distribution of a statistic is known as its 'standard error'. The standard error gives an idea of the average error that one would commit in using the value of the statistic in lieu of parameter. In the following equations, *n* denotes sample size, *N* denotes population size, σ denotes population standard deviation, *P* denotes population proportion such that ($P + Q = 1$).

i. Standard error of sample mean and sample proportion when sample is drawn with replacement.

- S.E. of Sample Mean $(\bar{x}) = \frac{\sigma}{\sqrt{n}}+$
- S.E. of Sample Proportion $(p) = \sqrt{\frac{PQ}{n}}$

ii. Standard error of sample mean and sample proportion when sample is drawn without replacement.

- S.E. of Sample Mean $(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
- S.E. of Sample Proportion $(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$

1.13. Mean (Expectation) and Standard Error of Sample Mean

Mean or expectation will be same for both random sampling with replacement and without replacement. Suppose a random sample of size 'n' is drawn from a given finite population of size 'N'. Let $x_1, x_2, x_3, \ldots \ldots, x_n$ denote the sample observations. The sample mean is

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

In simple sampling, each of the unit has the same probability distribution as the variable *x* in the population. Hence, $E(x_i) = \mu$ =Population Mean

$$\text{Var}(x_i) = E(x_i - \mu)^2 = \sigma^2 = \text{Population Variance}$$

Therefore, Sample Mean or Sample expectation $E(\bar{x}) = E\left[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\right]$

$$= \frac{1}{n}E(x_1 + x_2 + \cdots + x_n)$$

$$= \frac{1}{n}[E(x_1) + E(x_2) + \cdots + E(x_n)]$$

i.e., $E(\bar{x}) = \frac{1}{n}\sum E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu$

$$= \frac{1}{n}.n\mu = \mu$$

Variance of Sample Mean $\text{Var}(\bar{x}) = E(\bar{x} - \mu)^2$

$$= E\left[\sum\frac{(x_i - \mu)}{n}\right]^2 = \frac{1}{n^2}.E[\sum(x_i - \mu)]^2$$

$$= \frac{1}{n^2}.E\left[\sum_i (x_i - \mu)^2 + \sum\sum_{i \neq j}(x_i - \mu)(x_j - \mu)\right]$$

$$= \frac{1}{n^2} \cdot \left[\sum E(x_i - \mu)^2 + \sum \sum E\{(x_i - \mu)(x_j - \mu)\} \right]$$

But, $E(x_i - \mu)^2 = \text{Var}(x_i), E\{(x_i - \mu)(x_j - \mu)\} = \text{Cov}(x_i, x_j)$

$$\text{Var}(\bar{x}) = \frac{1}{n^2} \left[\sum_i \text{Var}(x_i) + \sum \sum_{i \neq j} \text{Cov}(x_i, x_j) \right] \qquad (2)$$

This result holds for the case simple random sampling with replacement and without replacement.

Case I, Simple Random Sampling with replacement (SRSWR)

Here x_i and x_j are independent.

$$\text{Var}(x_i) = \sigma^2 \quad \text{and} \quad \text{Cov}(x_i, x_j) = 0$$

Substituting these results in (2), we have

$$\text{Var}(\bar{x}) = \frac{1}{n^2} \left[\sum_i \sigma^2 + \sum \sum_{i \neq j} 0 \right] = \frac{1}{n^2} [n\sigma^2 + 0] = \frac{\sigma^2}{n}$$

Thus, in SRSWR, the Standard Error of sample mean is

$$\text{S. E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Case II, Simple Random Sampling without replacement (SRSWOR)

Here x_i and x_j are not independent.

$$\text{Var}(x_i) = \sigma^2 \quad \text{and} \quad \text{Cov}(x_i, x_j) = \frac{-\sigma^2}{N-1}$$

Substituting in (2)

$$\text{Var}(\bar{x}) = \frac{1}{n^2} \left[\sum_i \sigma^2 + \sum \sum_{i \neq j} \left(\frac{-\sigma^2}{N-1} \right) \right]$$

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

Because there are $n(n-1)$ possible pairs of values $i \neq j$.

$$\text{Var}(\bar{x}) = \frac{n\sigma^2}{n^2}\left[1 - \frac{n-1}{N-1}\right] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

$$\text{S. E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}\left(\frac{N-n}{N-1}\right)$$

1.14. Mean (Expectation) and Standard Error of Sample Proportion

Suppose in population of *N*. there are *Np* members with a particular character *A* and *Nq* members with the character not-*A*. Then *p* is the proportion of members in the population having the character *A*. Let a sample of size *n* be drawn from the population, and let *f* be the number of members in the sample having character *A*. To find the expectation and standard error of the sample proportion *f/n*, we adopt the following procedure.

We assign to the $\alpha - th$ member of the population the value X_α, which is equal to 1 if, this member possesses the character *A* and equal to 0 otherwise. Similarly, to the $i - th$member of the sample we assign the value x_i,, which is equal to 1 if this member possesses *A* and is equal to 0 otherwise.

In this way, we get a variable x, which has population mean

$$\frac{1}{N}\sum_\alpha X_\alpha = p$$

and population variance

$$\frac{1}{N}\sum_\alpha {X^2}_\alpha - p^2 = p - p^2 = pq \qquad \text{since } p + q = 1$$

The sample mean of the variable x, on the other hand, is

$$\frac{1}{n}\sum_i x_i = \frac{f}{n}$$

Hence we find, on replacing $\bar{x}$ by $\frac{f}{n}$, μ by *p* and σ^2 by *pq* in the expressions *E* ($\bar{x}$) and S.E.($\bar{x}$) given in preceding sections,

$E\left(\frac{f}{n}\right) = p$ [in case of random sampling with replacement]

$=p$ [in case of random sampling without replacement]

$\sigma_{f/n} = \sqrt{\frac{PQ}{n}}$ [In case of random sampling with replacement]

$= \sqrt{\frac{PQ}{n}} \sqrt{\frac{N-n}{N-1}}$ [in case of random sampling without replacement]

1.15. Probability Distribution derived from Normal Distribution

There are four fundamental distributions derived from Normal Distribution;

i. Standard Normal Distribution
ii. Chi-Square (χ^2) Distribution
iii. Student's *t* Distribution
iv. *F* Distribution

A. Standard Normal Distribution: If a random variable *x* is normally distributed with mean μ and standard deviation σ, then

$$Z = \frac{x - \mu}{\sigma} = \frac{\text{Normal Variable} - \text{Mean}}{\text{Standard Deviation}}$$

is called the ***Standard Normal Variate.*** The probability distribution of Z is called Standard Normal Distribution, and is defined by the probability density function

$$p(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} \quad ; (-\infty < Z < +\infty)$$

Characteristics of Standard Normal Distribution

1. Standard Normal Distribution is a special case of normal distribution with mean = 0 and standard deviation = 1.
2. It has no parameters.
3. The central moments are $\mu_2 = 1, \mu_3 = 0, \mu_4 = 3$. Also $\beta_1 = 0, \beta_2 = 3$; Skewness $(\gamma_1) = 0$, Kurtosis $(\gamma_2) = 0$.
4. The standard normal curve is symmetrical about the mean 0 and the two tails of the curve extend to infinity on either side of the mean. The points of infection are $Z = \pm 1$

B. Chi-Square (χ^2) Distribution: Let $Z_1, Z_2, \ldots, Z_n$ be *n* standard normal variables i.e., $Z_i \sim N(0,1), i = 1, 2, 3, \ldots\ldots, n$. Then, sum of squares of these variables, i.e., $\sum_i^n {Z_i}^2$ is said to have a χ^2 distribution with *n* degrees of freedom. The degree of freedom means the number of free or independent normal variable contained in χ^2 **.**

The probability density function of χ^2 distribution is given by

$$f(\chi^2) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-\chi^2/2} (\chi^2)^{n/2 - 1}, \qquad where\ 0 < \chi^2 < \infty$$

Characteristics of Chi-Square (χ^2) Distribution

1. It is continuous extending from 0 to ∞.
2. It is always positively skewed.
3. Its expectation is equal to its degrees of freedom and its variance is twice of its degrees of freedom.
4. It has additive property i.e., if Z_1, and Z_2 are two independent χ^2 variates with degrees of freedom n_1 and n_2 respectively, then $Z_1 + Z_2$ is also a χ^2 variable with degrees of freedom $n_1 + n_2$.

C. Student's *t* Distribution: Let *y* be a standard normal variable and *Y* a chi square with n degrees of freedom distributed independently of *y*, then the new variable

$$t = \frac{y}{\sqrt{Y/n}}$$

Is called a *t* variable with n degrees of freedom. It follows the probability density function as

$$f(t) = \frac{1}{n^{1/2}B(1/2,\ n/2)} \frac{1}{\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} \quad ; -\infty < t < \infty$$

Characteristics of Student's t Distribution

1. Mean = 0, Standard Deviation = $\sqrt{\frac{n}{n-2}}$, $(n > 2)$
2. The *t*-curve is symmetrical about 0, extending from -∞ to +∞. It has zero skewness and positive kurtosis, i.e., $\beta_1 = 0$, $\beta_2 > 3$
3. When the degrees of freedom *n* is large, the *t* distribution can be approximated by the standard normal distribution.

D. *F* Distribution: If Y_1 and Y_2 are two independent chi-square variates with n_1 and n_2 degrees of freedom respectively, then F statistic is defined by

$$F = \frac{Y_1/n_1}{Y_2/n_2}$$

In other words, *F* is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's *F* distribution with $(\boldsymbol{n_1}, \boldsymbol{n_2})$ degrees of freedom with probability function given by

$$f(F) = \frac{\left(n_1/n_2\right)^{n_1/2}}{B\left(n_1/2,\ n_2/2\right)} \frac{F^{(n_1-2)/2}}{\left(1 + n_1/n_2\, F\right)^{(n_1+n_2)/2}} \quad , \quad where\ 0 < F < \infty$$

Characteristics of F Distribution

1. Mean = $\frac{n_2}{n_2-2}$ Mode = $\frac{n_2(n_1-2)}{n_1(n_2+2)}$ Standard Deviation = $\left(\frac{n_2}{n_2-2}\right)\sqrt{\frac{2(n_1+n_2-2)}{n_1(n_2-4)}}$, provided they exist and are positive.
2. The *F*-curve is positively skewed, and starting from 0 extends to infinity.

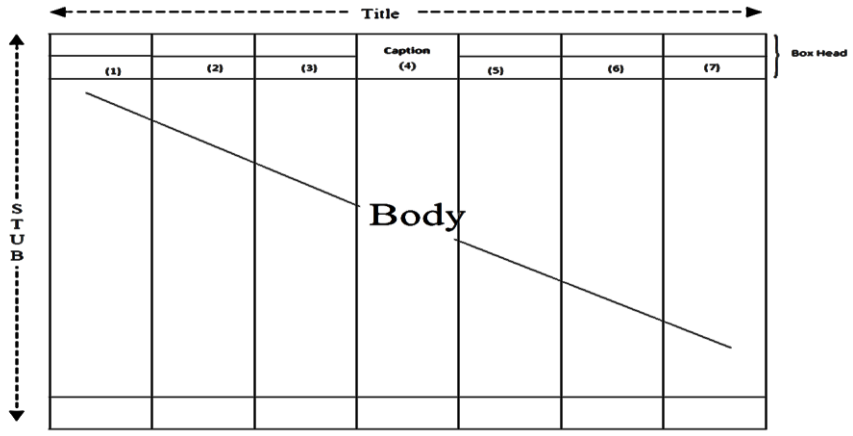
1.16. Data Presentation

After collecting the, the important work is to get the hidden information in the data out. This can be done by using appropriate technique to present these data. The technique can be tabular presentation or graphical presentation. Each technique has its own feature and can be used according to someone convenience. Therefore, the data are presented in two ways:

i. Tables
ii. Charts and Diagrams

1.16.1. Tabular presentation of Data: The tabular presentation of data is a logical and systematic organization of statistical data in rows and columns. It facilitates quick comparisons between variables and the error can be easily detected. In tabulation method repetition of data can be avoided and it reveals the characteristics of data clearly. A table constructed for presenting the data has the following parts:

i. **Title**: This gives the brief explanation of the contents and is shown at the top of the table.
ii. **Stub:** This is the extreme part of the table where information of rows are shown.
iii. **Caption and Box Head:** The upper part of the table, which shows the description of columns and sub columns, is called Caption. The row of the upper part, including caption, units of measurement and column number, if any, is called box-head.
iv. **Body:** This part of the table shows the figures.
v. **Footnote:** In this part below the body, where the source of data and explanations are shown, if any.



SOURCE……

FOOT NOTE……

Figure 1.1: Different Parts of Table

1.16.2. Diagrammatic presentation for Ungrouped Data

The charts and diagrams are also use to present data. The positive point of diagrammatic presentation is that we can get an idea about the nature or trend of the data by just looking at it. Charts and diagrams are also useful to find whether any relation exists between two or more set of data. They But charts or diagrams unlike tables do not show details of data and require much time to construct. There are mainly three types of charts and diagrams:

i. Line Diagram
ii. Bar Diagram
iii. Pie Diagram

(i) Line Diagram: The most common method of presenting data is line diagram. Data presentation in the form of line diagrams are mostly used in business and commerce. Mostly, the time series data are represented by line diagrams. In case of line diagram two mutually perpendicular straight lines are taken as axes. On the horizontal axis units of X are represented and on the vertical axis units of Y are represented. The intersection of these two lines is taken as the origin. The given data are represented as points on the graph paper. The locus of all such points joined either by curves or by pieces of straight lines gives the line diagram.

Two types of line diagrams are used, natural scale and ratio scale. In the natural scale equal distances represent equal amounts of change. But in ratio scale equal distances represent equal ratios. Below we provide an example of line diagram.

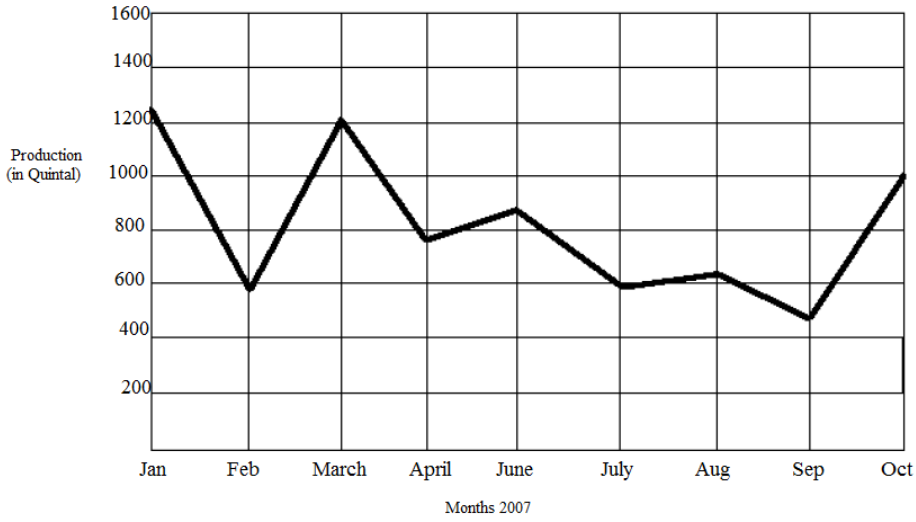


Figure 1.2 Line diagram showing production of a firm against months of 2007

(ii) Bar Diagram: Bar diagrams may be used to present data on time series or spatial data. In this method rectangles or bars of equal width are drawn for each of the items of the series and the length of the bar represents the value of the concerned variable. Generally, vertical bars are taken for time series data while horizontal bars are taken for spatial data. The bar starts from a common base and the bars should be equally spaced.

Example 1. *Represent the following data by a line and bar diagrams showing the difference between proceeds and costs.* [I.C.W.A., Jan 1966]

Year	*Total Proceeds*	*Total Costs*
1950	22.0	19.5
1951	27.3	21.7
1952	28.2	30.0
1953	30.3	25.6
1954	32.7	26.1
1955	33.3	34.2

Solution: Let us first draw the line diagram as:

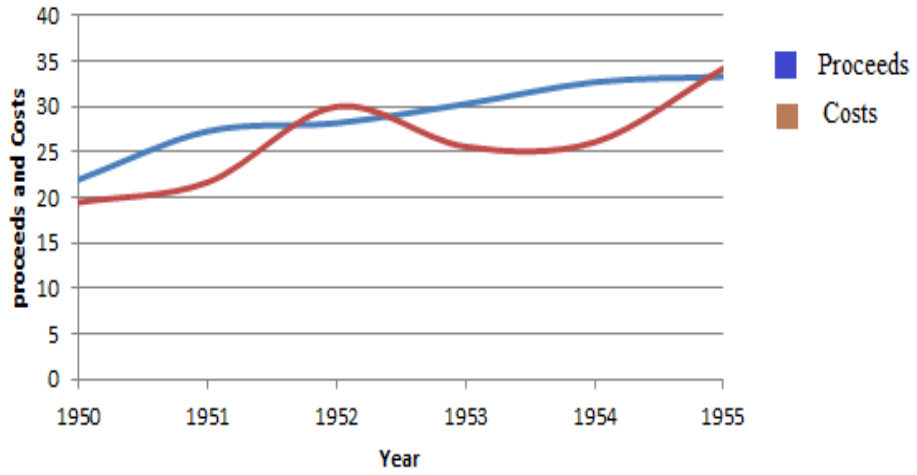


Figure 1.3: Grouped Line Diagram

And the bar diagram is

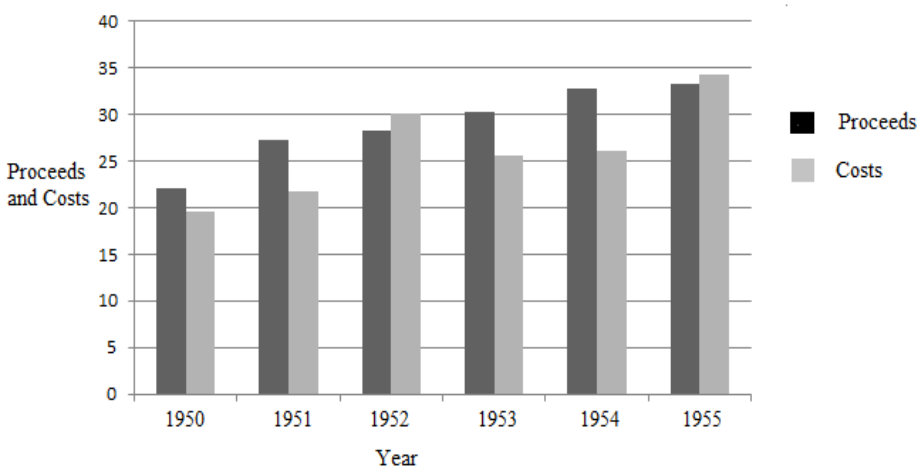


Figure 1.4 Grouped bar Diagram

(iii) Pie Diagram: When the comparison is made between the relative sizes of the component parts of a whole. Hence, a circle is used and the area of the circle is taken as 100%. Thus, the area is divided proportionately among the different components by straight lines drawn from the center to the circumference. While drawing a pie diagram it is necessary to express the value of each category as a percentage of the total. Since the circle represents the total and to represent each category in that circle, we have to multiply the percentage of each category by 3.6 degrees, so that sum of each category becomes 360 degrees. The diagram can be drawn with the help of a compass and a protractor.

Example2. *Construct a pie chart for the following data of principal exporting countries of cotton (in 1,000 bales) for the year 1955-56.* [C.U., M.Com. 1974]

U.S.A.	*India*	*Egypt*	*Brazil*	*Argentina*
6367	2999	1688	650	202

Solution: Calculation for Pie Chart

Country	*Exports*	*Percent of Total*	*Angle at the center of pie chart: col 3 X 3.6*
U.S.A.	6367	$(6367 \div 11906) \times 100 = 54$	194.4
India	2999	$(2999 \div 11906) \times 100 = 25$	90.0
Egypt	1688	$(1688 \div 11906) \times 100 = 14$	50.4
Brazil	650	$(650 \div 11906) \times 100 = 5$	18.0
Argentina	202	$(202 \div 11906) \times 100 = 2$	7.2
Total	11906	100	360.0

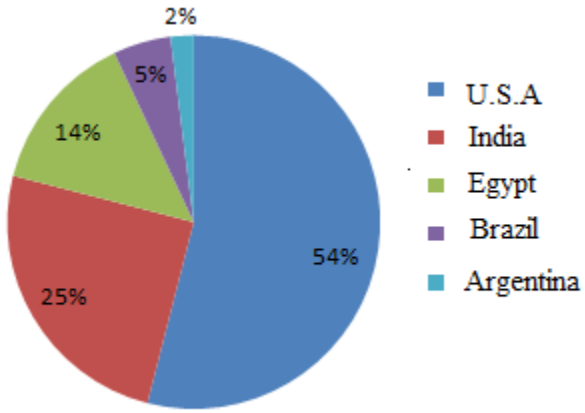


Figure 1.5 Pie Chart Showing Principal Exporting Countries of Cotton

1.16.3. Simple (or Ungrouped) Frequency Data and Grouped Data or Frequency Distribution

Frequency of a variable is defined as the number of times it occurs in the given data. Suppose we have a large where observations are repeated many times, then we take the number of time an observation occurred in the group. This number is known as frequency. The frequency distribution is of two types; Simple Frequency Distribution and Grouped Frequency Distribution. In case simple frequency distribution shows the values of the variable individually whereas the grouped frequency distribution shows the value of a variable in the group or interval. Let us suppose that the variable x takes the values or observations on x as $x_1, x_2, \ldots\ldots, x_n$ and the corresponding frequencies are $f_1, f_2, \ldots\ldots, f_n$ Then frequency of x_i is generally denoted by f_i.Let us consider the table 1.1.where the frequency distribution of marks obtained by 50 students is shown.

Marks Obtained by Students (x_i)	Number of Students (f_i)
30	8
45	9
60	7
53	6
69	3
75	2
43	4
50	11

Table 1.1: Simple Frequency Distribution of Marks

The table 1.1 shows 8 students scored 30 marks, 9 students scored 45 and so on. Thus, for each of variable x_i we have corresponding frequencies. Since the observation is small it is easy to grasp the nature of the marks obtained by the students. But, when the data are large and to grasp the hidden facts grouped frequency distribution is used. In the following table (Table 1.2) we shall consider a grouped frequency distribution and some basic terminologies associated with it.

- **Basic Terminology:**

(i) Class or Class Interval: When a large number of observations varying in a wide range are available, they are usually classified into several groups according to the size of the values. Each of these groups defined by an interval is called class interval or simply class. In the table 1.2, 15-19, 20-24 and so on are the class intervals. There are total 6 classes in the table 1.2.

(ii) Class Frequency: The number of observation falling under each class is called its class frequency or simply frequency. The sum of these frequencies is known as total frequency. In the table 1.2, 200 is the total frequency.

(iii) Cumulative Frequency: Cumulative frequency corresponding to a specified value of a variable or a class (in case of grouped frequency distribution) is the number of observations smaller (or greater) than that value or

class. The number of observation up to a given value (or class) is called less-than type cumulative frequency distribution, whereas the number of observations greater than a value (or class) is called more-than type cumulative frequency distribution. The column 3 of the table 1.2 shows cumulative frequency (both less than and more than).

(iv) Class Limits: The upper and the lower limit of the class interval is class limit. The two numbers used to specify the limits of a class interval for tallying the original observations.

(v) Class Boundaries: For a continuous variable measurement, all data are recorded nearest to a certain unit. The class interval 15-19actually includes the value from 14.5-19.5. These are in fact the real class limit of a class limit. Thus, the extreme values (observations) of a variable, which could ever be included in a class interval, are called class boundaries.

(vi) Mid-Point of Class Interval: The value exactly at the middle of a class interval is called class mark or mid-value. It is used as the representative value of the class interval. Thus, Mid-point of Class interval = (Lower class boundary +Upper class boundary)/2.

(vii) Width of a Class: Width of class is defined as the difference between the upper and lower class boundaries. Thus, Width of a Class = (upper class boundary - lower class boundary).

(viii) Relative Frequency: The relative frequency of a class is the share of that class in total frequency. Thus, Relative Frequency = (Class frequency / Total frequency).

(ix) Frequency Density: Frequency density of a class is its frequency per unit width. Thus, Frequency density = (Class frequency / Width of the class).

Class Interval	*Class Frequency*	*Cumulative Frequency*		*Class Limits*		*Class Boundaries*		*Class Marks*	*Width of the class*	*Frequency Distribution*	*Relative Frequency*
		Less Than	*More Than*	*Lower*	*Upper*	*Lower*	*Upper*				
15-19	37	37	200	15	19	14.5	19.5	17	5	7.4	0.185
20-24	81	118	163	20	24	19.5	24.5	22	5	16.2	0.405
25-29	43	161	82	25	29	24.5	29.5	27	5	8.6	0.215
30-34	24	185	39	30	34	29.5	34.5	32	5	4.8	0.120
35-44	9	194	15	35	44	34.5	44.5	39.5	10	0.9	0.045
45-59	6	200	6	45	59	44.5	59.5	52	15	0.4	0.030

Total	200	-	-	-	-	-	-	-	-	-	1.00

Table 1.2: Grouped Frequency Distribution of Ages

1.16.4. Diagrammatic Presentation of Frequency Distribution

The diagrams commonly used to depict the frequency distribution are;

i. Histogram
ii. Frequency Polygon
iii. Ogive (or Cumulative Frequency Polygon)

(i) Histogram: To present the data of a frequency distribution, histogram is more preferred. It consists of a set of adjacent rectangles drawn on the horizontal base line, with areas proportional to the class frequencies. The base of each rectangle measures the class width whereas the height measures the frequency density. The height of the rectangles will be proportional to the class frequencies if the classes have the equal widths. If not, the rectangles will be of unequal width if the classes are of unequal width and therefore the heights must be proportional to the frequency densities.

Though the bar diagram and the histogram looks same, in bar diagram the rectangles are spaced equally whereas the consecutive rectangles in histogram have no space in between.

(ii) Frequency Polygon: The frequency polygon is derived from the histogram by joining the mid points of the tops of the consecutive rectangles. It is used in cases when all the classes have a common width. The two end points of a frequency polygon are joined to the base line at the mid values of the empty classes at the end of the frequency distribution.

(iii) Ogive (or Cumulative Frequency Polygon): The ogive the diagrammatic presentation of cumulative frequencies and hence is also called Cumulative Frequency Polygon. When the cumulative frequencies are plotted against the class boundaries and the successive points are joined by a line, the line diagram thus obtained is known as Ogive. There are two types of ogive; less than and more than type. The less than type ogive starts from the lowest class boundary on the horizontal axis to the highest class boundary. Ogives are used to determine median, quartiles, deciles and percentiles.

Example 3: *Following data were obtained from a survey on the value of annual sales of 534 firms. Draw the histogram and the frequency polygon and ogive from the data.* [C.U., B.A. (Econ) 1963]

Values of Sales	*Number of firms*
0-500	3
500-1000	42
1000-1500	63
1500-2000	105
2000-2500	120
2500-3000	99
3000-3500	51
3500-4000	47
4000-4500	4

Solution: Since here the interval is defined by the class boundaries, therefore it is relatively easier to draw histogram and frequency polygon.

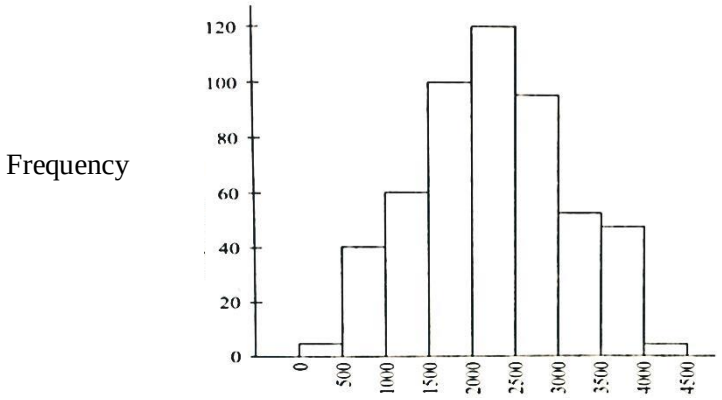


Figure 1.6. Histogram

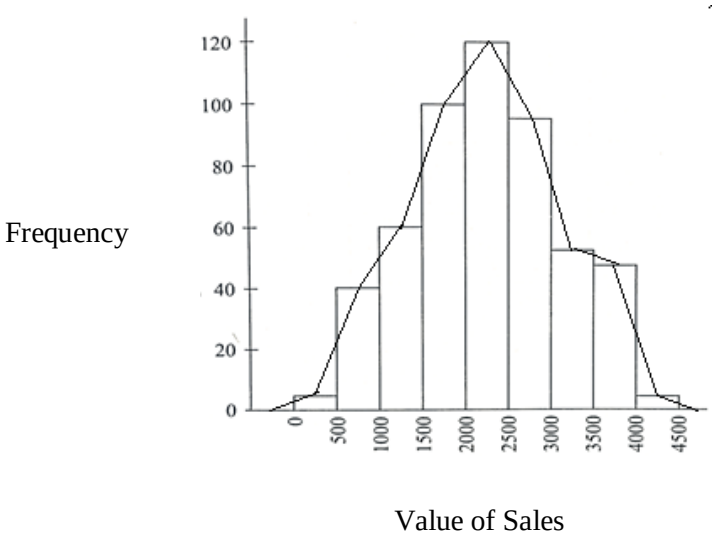


Figure 1.7. Frequency Polygon

In order to draw the ogive we have to construct the cumulative frequency distribution from the above data. It is done in the next table.

Class Boundary	*Cumulative Frequency (Less-than)*
0	0
500	3
1000	45
1500	108
2000	213
2500	333
3000	432
3500	483
4000	530
4500	534 = N

We plot the above data to get the following ogive.

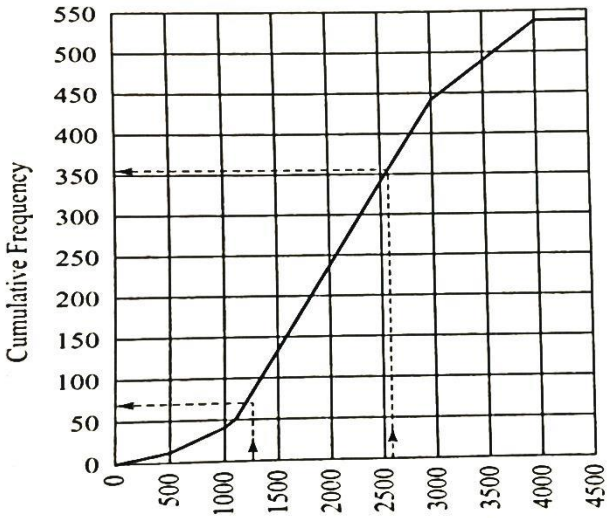


Figure 1.8. Ogive (Less than)

Exercises

1. Define population and sample. What are the advantages of sampling?

2. Define parameter and statistic.

3. What are simple random sampling and stratified random sampling?

4. Differentiate systematic sampling and cluster sampling.

5. Explain standard error. State standard error of sample mean and sample proportion.

6. Prepare a blank table showing the distribution of population according to gender and four religions in five age groups in seven different cities. [C.U., B.A. Econ 1972]

7. The following data shows the numbers of agricultural and non-agricultural workers for the year 1860-1950. Graph the data using (a) Line diagrams and (b) Bar diagrams. [C.U., M.Com. 1967]

Year	*1860*	*1870*	*1880*	*1890*	*1900*	*1910*	*1920*	*1930*	*1940*	*1950*
Agricultural Workers (in million)	6.2	6.9	8.6	9.9	10.9	11.6	11.4	10.5	8.8	6.8
Non-Agricultural Workers (in million)	4.3	6.1	8.8	13.4	18.2	25.4	31.0	38.4	42.9	52.2

8. Draw a pie diagram to represent the following data of proposed expenditure by a State Government for the year 1997-98. [D.U., B.Com (Pass). 1997]

Items	*Agriculture and Rural Development*	*Industries and Urban Development*	*Health and Education*	*Miscellaneous*
Proposed expenditure (In Million Rs.)	4,200	1,500	1,000	500

9. Draw a histogram, frequency polygon and ogive from the following data. [C.S., (Foundation). June 2000]

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	4	6	7	14	16	14	8	6	5

10. Draw a less than ogive from the following data. [D.U., B.Com (Hons). 1997]

Weekly income (Rs.)	12,000	11,000	10,000	8,000	6,000	4,000	3,000	2,000	1,000
No. of Families	0	6	14	26	42	54	62	70	80

UNIT II **MEASURES OF CENTRAL TENDENCY**

2.1 Objectives: - This chapter enables us to know how a single value can be a representative of a group of values and how these values of central tendencies used to denote economic factors in an economy.

2.2 Introduction: In the previous chapter, we learn how to collect data, its classification and tabulation in a useful form. Yet, this is insufficient to describe the nature of a distribution. Hence, further consideration is required to compare two or more distribution. For that we may reduce the whole distribution to one number which represents the whole distribution.

The word 'Average' is defined as a single value which can be considered as a typical representative of a set of observations and round the observations. If the observations are arranged according to magnitudes, this single value tends to lie in center thus averages are called measure of central tendency. Since, it occupies a central position, some observations are larger and some are smaller than it.

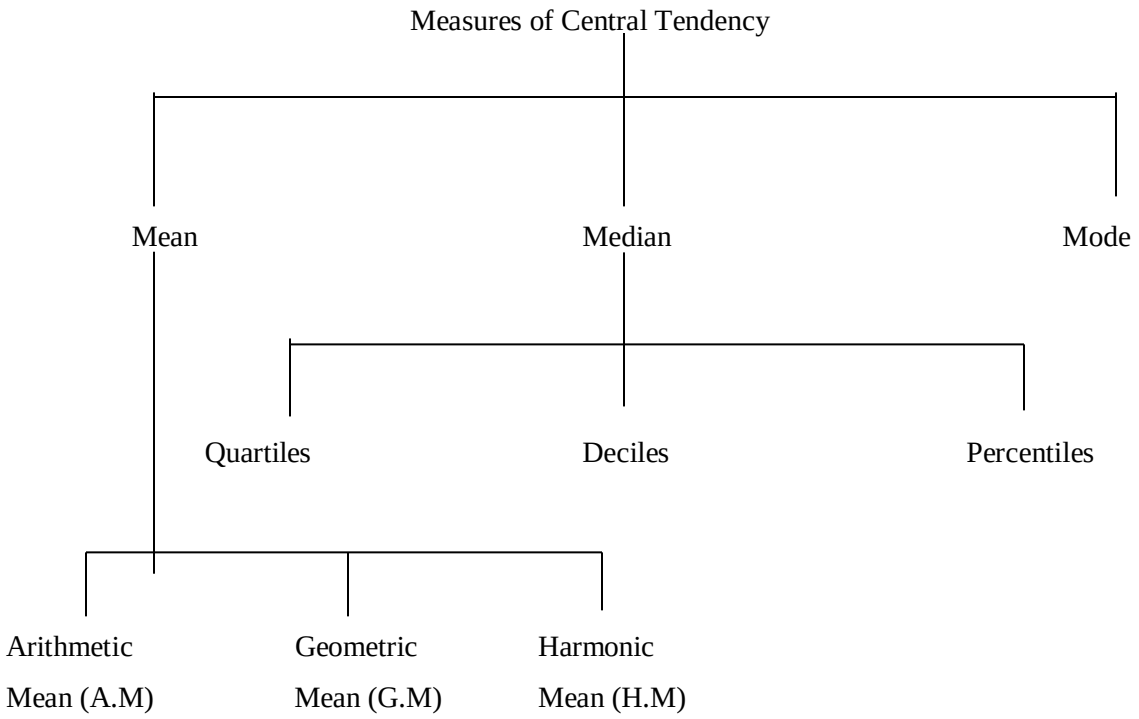
2.3. Requisites of a Good Average

1. *Based on all observations:* A good measure of average should use all the observations available on data and there should not be any loss of information. Only then it will provide a good result.
2. *Should not be affected much by extreme values:* The extreme values either are very small or very large observation. Thus, a good measure of average should not be unduly affected by these small or large observations.
3. *Should be easy to understand and calculate:* A good average is a one which is easy to understand and simple to calculate. While calculating an average a heavy arithmetical procedure is followed, it will be not easy for a person to understand. Thus, calculations involved while measuring average should be as simple as possible.
4. *Should be rigidly defined:* The definition should be clear and unambiguous so that it leads to one and only one interpretation by different people.
5. *Should be suitable for further mathematical treatment:* Methods of central tendency are used in many other techniques of statistical analysis like measures of dispersion, correlation etc.
6. *Should not be affected by sampling variations:* A good average is a one which is least affected by sampling fluctuations. If a few samples are taken from same population, the average should be such as has the least variation in values in the values derived in the individual samples. The results obtained will be considered to be the true representative of the population in this case.

2.4. Measures of Central Tendency

There are three measures of central tendency:

i) Mean, ii) Median and iii) Mode and these measures are further divided into sub categories. They are presented in a form of a tree figure.



Let us now discuss these measures one by one.

2.5. Mean

2.5.1 Arithmetic Mean

The arithmetic mean of a variable is obtained by dividing the sum of its given values by their number. Depending on whether the data are grouped or ungrouped arithmetic mean may be of two types. First, simple arithmetic mean for ungrouped data and second, weighted arithmetic mean for grouped (frequency type) data. For a ungrouped data, if the variable is denoted by x and n values of x are given, viz. $x_1, x_2, \dots\dots\dots, x_n$, then the arithmetic mean of x is

$$\bar{x} = \sum_{i=1}^{n} x_i / n \qquad where\ i = 1 \dots\dots n$$

$$= \frac{x_1 + x_2 + \cdots + x_n}{n}$$

If the variable x takes the values $x_1, x_2, \dots\dots\dots, x_n$, with frequencies f_1, f_2, $\dots f_n$ then Weighted arithmetic mean

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

Example 1: *Given the following data calculate the simple and weighted arithmetic mean*

x :	10	20	30	40	50	60
f :	6	4	6	12	8	4

Solution:

Calculations for A.M.

X	*F*	*fx*
10	6	60
20	4	80
30	6	180
40	12	480
50	8	400
60	4	240
$\sum x = 210$	$\sum f = 40$	$\sum f\,x = 1440$

$$\therefore Simpl\ A.M = \frac{\sum x}{n} = \frac{210}{6} = 35$$

And $Grouped\ A.M = \frac{\sum fx}{\sum f} = \frac{1440}{40} = 36.$

Example 2: *The following table gives marks of two candidates; (1) find the weighted average mark of each candidate. (2) By what figure would the 2nd candidate have had to increase his marks in subject B, all other marks remaining same, in order that both the candidates have the same place?*

Subject	*Weight*	*Marks of*	
		1st candidate	*2nd candidate*
A	1	70	80
B	2	65	64
C	3	58	56

D	4	63	60

Solution: *Table 2.2.* Calculations for Weighted Arithmetic Mean

Subject	*Marks of*		*Weight*	*fx*	*fy*
	1st candidate (x)	*2nd candidate (y)*	*F*		
A	70	80	1	70	80
B	65	64	2	130	128
C	58	56	3	174	168
D	63	60	4	252	240
Total	-	-	10	626	616

(1) Weighted arithmetic means of marks are:

(a) 1st candidate = $\frac{\sum fx}{\sum f} = \frac{626}{10} = 62.6$

(b) 2nd candidate = $\frac{\sum fy}{\sum f} = \frac{616}{10} = 61.6$

(2) Let the required increase in marks be K. then the marks of 2nd candidate in subject B would be (64+K). Hence, in table 2.2, the number 64 against subject B in the 3rd column would be 64+K; and consequently in the last column 128 should be replaced by 2×(64+K) =128+2K. The total of the last column would then be 616+2K. in order that both candidate have the same place, their total marks should be equal. Hence

616+2K = Total marks of 1st candidate = 626

Or, 2K = 626 -616 = 10

∴ K = 10/2 = 5.

2.5.1.1. Properties of Arithmetic Mean

i) The algebraic sum of the deviations of the values from their arithmetic mean is zero.

Proof. (a) *(Simple Arithmetic Mean*) Let $x_1, x_2, \dots\dots\dots, x_n$ be a series of number, and $\bar{x}$ their arithmetic mean. The deviations of the number from their A.M are

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots\dots., (x_n - \bar{x})$$

Then, the algebraic sum of these deviations is

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

$$= (x_1 + x_2 + \cdots + x_n) - \left(\bar{x} + \bar{x} + \cdots + \bar{x}(n\ times)\right)$$

$$= \sum x - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

(b)*Weighted Arithmetic Mean*

$$\sum_{i=1}^{n} f_i\,(x_i - \bar{x}) = f_1(x_1 - \bar{x}) + f_2(x_2 - \bar{x}) + \ldots\ldots\ldots + f_n(x_n - \bar{x})$$

$$= f_1x_1 + f_2x_2 + \cdots + f_nx_n - (f_1\bar{x} + f_2\bar{x} + \cdots + f_n\bar{x})$$

$$= \sum f_i\bar{x} - (f_1 + f_2 + \cdots + f_2)\bar{x}$$

$$= \sum f_i\bar{x} - = \sum f_i.\bar{x} = 0, \qquad (as\ \bar{x} = \frac{\sum f\bar{x}}{\sum f})$$

ii) The sum of squares of deviations of a set of observations has the smallest value, when deviations are taken from their A.M.

Proof: We can write

$$x_i - A = (x_i - \bar{x}) + (\bar{x} - A), \qquad where\ x_1 are\ n\ observations.$$

$\bar{x}$ is actual mean, A is any arbitrary constant.

Now $\sum(x_i - A) = \sum(x_i - \bar{x}) + \sum(\bar{x} - A)$ $[taking\ \sum both\ sides]$

$\sum(x_i - A)^2 = \sum(x_i - \bar{x})^2 + \sum(\bar{x} - A)^2 + 2\sum(x_i - \bar{x})(\bar{x} - A)$ $[squaring\ both\ sides]$

$= \sum(x_i - \bar{x})^2 + n((\bar{x} - A)^2 + 2(\bar{x} - A)\sum(x_i - \bar{x}),$ $as\ ((\bar{x} - A)\ is\ constant.$

$= \sum(x_i - \bar{x})^2 + n(\bar{x} - A)^2,$ $[as \sum(x_i - \bar{x}) = 0, by\ prop.(1)]$

The right side is the sum of two positive numbers. We are to find the value of A for which $\sum(x_i - A)^2$ is minimum. Now $\sum(x_i - A)^2$ will be minimum, if $n(\bar{x} - A)^2$ is minimum, i.e., if it is zero. Now if

$n(\bar{x} - A)^2 = 0$, then $\bar{x} - A = 0\ (as\ n \neq 0) or, A = \bar{x}$.

$\sum(x_i - A)^2$ will be minimum, if $A = \bar{x}$.

iii) If two variables x and y are so related that $y_i = x_i - c$, $(i = 1, 2, \ldots\ldots, n)$ where c is a constant, then $\bar{x} = c + \bar{y}$.

Proof: Since $y_i = x_i - c$, therefore $x_i = c + x_i$

Multiplying both sides by f_i and then summing over all values of i=1, 2,…., n we have

$$\sum_{1}^{n} f_i x_i = \sum_{i}^{n} f_i(c + y_i)$$

$$= \sum(f_i c + f_i y_i)$$

$$= \sum f_i c + \sum f_i y_i$$

$$= c\sum f_i + \sum f_i y_i$$

$$= cN + \sum f_i y_i \qquad since \sum f_i = N$$

Hence, $\bar{x} = \frac{1}{N}\sum f_i x_i = \frac{1}{N}(cN + \sum f_i y_i) = c + \frac{1}{N}\sum f_i y_i = c + \bar{y}$

The significance of this result is that if $y_1, y_2, \ldots\ldots\ldots, y_n$ are the deviations of $x_1, x_2, \ldots\ldots\ldots, x_n$ from an arbitrary constant *c*, then

Mean of x = c + Mean of y.

iv) If two variables x and y are so related that $\boldsymbol{y_i = \frac{x_i - c,}{d}(i = 1, 2, \ldots\ldots, n)}$ ***where c and d are constants, then*** $\boldsymbol{\bar{x} = c + \overline{dy}}$.

Proof: [Note: *c* and *d* are known as origin and scale respectively of the values y_i.]

Since $y_i = \frac{x_i - c,}{d}, we\ have\ dy_i = x_i - c$

Or, $x_i = c + dy_i$

Multiplying both sides by f_i , we get $f_i x_i = f_i(c + dy_i)$

Now, summing over all values of *i*=1, 2, ……., *n*

$$\sum f_i x_i = \sum f_i(c + dy_i) = \sum(f_i c + d f_i y_i)$$

$$= \sum f_i c + \sum d f_i y_i$$

$$= c\sum f_i + d\sum f_i y_i$$

$$= cN + d\sum f_i y_i$$

Hence, $\bar{x} = \frac{1}{N}\sum f_i x_i = \frac{1}{N}(cN + d\sum f_i y_i) = c + d\left(\frac{1}{N}\sum f_i y_i\right) = c + d\bar{y}$

This means that $y_1, y_2, \ldots\ldots\ldots, y_n$ be the deviations of $x_1, x_2, \ldots\ldots\ldots, x_n$ from an arbitrary constant *c*, in units of another constant *d*, then *Mean of x = c + d (Mean of y).*

Examples 3: *Compute the average weekly wages of the 65 employees working in a factory from the frequency table, using the coding method (i.e. transforming x to a new variate):*

x:	55	65	75	85	95	105	115
y :	8	10	16	14	10	5	2

Solution: We introduce the new variable $y = \frac{x-85}{10}$, where *c=85* (one of the given values of *x*, preferably near the middle) and *d* = 10 (common difference)

Calculations for Mean:

x	*F*	$y = \frac{x-85}{10}$	*fy*
55	8	-3	-24
65	10	-2	-20
75	16	-1	-16
85	14	0	0
95	10	1	10
105	5	2	10
115	2	3	6
Total	65	-	-34

$$\bar{x} = c + \overline{dy}$$

$$= 85 + 10\bar{y}$$

$$= 85 + 10\left(\frac{-34}{65}\right)$$

$$= 85 - 5.23 = 79.77$$

- **Calculation of Mean from Grouped Data** (Continuous Series)

Example 4: *Calculate the mean from the following frequency distribution.*

Wages (Rs.):	12.5-17.5	17.5-22.5	22.5-27.5	27.5-32.5	32.5-37.5	37.5-42.5
No. of workers:	2	22	10	14	3	4

42.5-47.5	47.5-52.5	52.5-57.5	Total
6	1	1	63

Solution: Take the mid-values of class intervals as *x* and follow the steps of coding method.

Class interval	*f*	*Mid-value* *x*	$y = \frac{x - 35}{5}$	*fy*
12.5-17.5	2	15	-4	-8
17.5-22.5	22	20	-3	-66
22.5-27.5	10	25	-2	-20
27.5-32.5	14	30	-1	-14
32.5-37.5	3	35	0	0
37.5-42.5	4	40	1	4

42.5-47.5	6	45	2	12
47.5-52.5	1	50	3	3
52.5-57.5	1	55	4	4
Total	63	-	-	-85

Here, $y = \frac{x-35}{5}$, where c = 35 and d = 5.

$$\bar{x} = c + \overline{dy}$$

$$= 35 + 5\bar{y}$$

$$= 35 + 5\left(\frac{-85}{63}\right)$$

$$= 35 - 6.75$$

$$= 28.25\ (Rs)$$

2.5.1.2. Mean of Combined Group

If two groups contains n_1 and n_2 observations with means x_1 and x_2 respectively, then the mean $(\bar{x})$ of the combined group of $n_1 + n_2$observations is given by the relation

$$N\bar{x} = n_1\overline{x_1} + n_2\overline{x_2}$$

Where $N = n_1 + n_2$

Example 5: *The mean weight of 150 students (boys and girls) in a class 60 kg. The mean weight of boy-students is 70 kg and that of girl-student is 55 kg. Find the number of boys and girls in that class.*

Solution: Let number of boy-student = n_1, number of girl-student = n_2

Now, $n_1 + n_2 = 150$, again $\bar{x}$ =60 kg, $\bar{x}_1 = 70$ kg and $\overline{x_2} = 55$ kg

From the formula $\bar{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$, $we\ find$ $\quad 60 = \frac{n_1 \times 70 + n_2 \times 55}{150}$

$$or, \quad 70n_1 + 55n_2 = 60 \times 150 = 9000 \quad or, \quad 70(150 - n_2) + 55n_2 = 9000$$

$$or, \quad 10500 - 70n_2 + 55n_2 = 9000 \quad or, \quad -15n_2 = -1500$$

$$or, \quad n_2 = 100 \quad \therefore n_1 = 150 - n_2 = 150 - 100 = 50$$

$\therefore$ Required no. of boy-student = 100, number of girl-students = 50.

2.5.1.3. Merits and Demerits of A.M.:

Merits:

1. It is rigidly defined. Its value is always defined.
2. It is easy to calculate and easy to understand. Hence it is very popular.
3. It is based on all observations; so that it becomes a good representative.
4. It can be easily used for comparison.

Demerits:

1. It is affected by extreme values. For example the A.M. of 10, 15, 25 and 500 is 137.5. Now observe first three values whose A.M. is 16.67 (approx.). Due to extreme value 500 the A.M. of the four numbers is raised to 137.5. In such a case A.M. is not a good representative of the given data.
2. It is a value which may not be present in the given data.
3. Many a times it gives absurd results like 4.4 children per family.
4. It is not possible to take the averages of ratios and percentages.
5. We cannot calculate it when open-end class intervals are present in the data.

2.5.2. GEOMETRIC MEAN (G.M)

Geometric mean of a group of n observations is the $n - th$ root of their product. It is defined only when all observations have the same sign and none of them is zero.

1) **Simple Geometric Mean:** Given n observations $x_1, x_2, \ldots\ldots\ldots, x_n$,

$$\text{G.M} = \sqrt[n]{x_1 \times x_2 \times \ldots\ldots \times x_n} \quad \text{or, } (x_1 \times x_2 \times \ldots\ldots \times x_n)^{\frac{1}{n}}$$

Now, taking logarithms on both sides we find,

$$\log G.M = \frac{1}{n}\log(x_1 \times x_2 \times \ldots\ldots \times x_n)$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log x_n) = \frac{1}{n}\sum \log x_i$$

$$\therefore G.M = antilog\left(\frac{1}{n}\sum \log x_i\right)$$

2) Weighted Geometric Mean: If n positive values of a variate $x_1, x_2, \ldots\ldots\ldots, x_n$, are taken respectively $f_1, f_2, \ldots\ldots f_n$, then their weighted G.M. is given by

$$\left({x_1}^{f_1} \times {x_2}^{f_2} \times \ldots\ldots \times {x_n}^{f_n}\right)^{1/N}$$

Where N = $f_1 + f_2 + \ldots\ldots + f_n$ $\quad i.e., N = \sum f_i$

Taking log, we have $\log G.M = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \cdots + f_n \log x_n) = \frac{1}{N}\sum f_i \log x_i$

$$G.M = antilog\left(\frac{1}{N}\sum f_i \log x_i\right)$$

2.5.2.1. Properties of G.M.

1. The product of n observations is equal to the n-th power of their G.M.

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

$$G.M^n = (x_1 \times x_2 \times \ldots\ldots \times x_n)$$

2. The logarithm of G.M. of a set of observations is equal to the A.M. of their logarithms.

i.e., $\log G.M = \frac{1}{n}\sum \log x_i$ $\quad and \quad \log G.M = \frac{1}{N}\sum f_i \log x_i$

3. If $G_1, G_2, \ldots$ are the geometric mean of different groups having observations $n_1, n_2, \ldots\ldots$ respectively, then G.M (G) of the composite group is given by

$$G = \sqrt[N]{{G_1}^{n_1} \times {G_2}^{n_2} \times \ldots\ldots\ldots}$$

$$\log G = \frac{1}{N}\sum n_i(\log G_i)$$

Example 6. *Find the G.M. of 111, 171, 191, and 212 having weights by 3, 2, 4, and 5 respectively.*

Solution: Calculations for G.M.

x	*F*	*Log x*	*f log x*
111	3	2.0453	6.1359

171	2	2.2330	4.4660
191	4	2.2810	9.1240
212	5	2.3263	11.6315
TOTAL	14		31.3574

$$\log G.M. = \frac{\sum f \log x}{\sum f} = \frac{31.3574}{14} = 2.2398 \ \therefore G.M = antilog\ 2.2398 = 173.7$$

Example 7. *A machine depreciates 25% of its value during the first two years, 10% for the next three years and 2% in value for the next five years, depreciation being calculated on the diminishing value. If the value of the machine is Rs.5000, find the depreciated value after 10 years.*

Solution: In the calculation of depreciated values, the Compound interest law should be applied, with *i* negative.

$$A = P(1-i_1)^{n_1}\ (1-i_2)^{n_2}\ (1-i_3)^{n_3}$$

$$= 5000\left(1-\frac{25}{100}\right)^2\ \left(1-\frac{10}{100}\right)^3\ \left(1-\frac{2}{100}\right)^5$$

$$= 5000 \times (0.75)^2 \times (0.90)^3 \times (0.98)^5$$

$$\log A = \log 5000 + 2(\log 0.75) + 3(\log 0.90) + 5(\log 0.98)$$

$$= 3.6990 + 2 \times \bar{1}.8751 + 3 \times \bar{1}.9542 + 5 \times \bar{1}.9912$$

$$= 3.6990 + 2(-1 + .8751) + 3(-1 + .9542) + 5(-1 + .9912)$$

$$= 3.6990 + (-2 + 1.7502) + (-3 + 2.8626) + (-5 + 4.9560)$$

$$= 13.2678 - 10 = 3.2678$$

$$\therefore A = antilog\ 3.2678 = 1853\ Rs.$$

2.5.2.2. Merits and Demerits of G.M.

Merits:

1. It is not influenced by the extreme items to the same extent as mean.
2. It is rigidly defined and its value is a precise figure.
3. It is based on all observations and capable of further algebraic treatment.
4. It is useful in calculating index number.

Demerits:

1. It is neither easy to calculate nor is simple to understand.
2. If any value of a set of observations is zero, the G.M. would be zero, and it cannot be determined.
3. If again any value becomes negative, G.M. becomes imaginary.

2.5.3. HARMONIC MEAN (H.M)

2.3.1. Harmonic Mean of a set of observations is the reciprocal of the arithmetic mean of their reciprocal. Like G.M., H.M is defined only when no observations is zero.

1. Simple H.M. $= \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$

2. Weighted H.M.$= \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{fn}{x_n}} = \frac{N}{\sum\left(\frac{f_i}{x_i}\right)}$

Examples 8: *In a certain office, a letter is typed by A in 4 minutes. The same letter is typed by B, C and D in 5, 6 and 10 minutes respectively. What is the average time taken in completing one letter? How, many letters do you expect to be typed in one day of 8 working hours?*

Solution: The average time (minutes) taken by each of A, B, C and D in completing one letter is the harmonic mean of 4, 5, 6 and 10 given by:

$$= \frac{4}{\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{10}} = \frac{4}{\left(\frac{15+12+10+6}{60}\right)} = \frac{240}{43} = 5.5814 \text{ minutes per letter.}$$

Hence, expected number of letters typed by each of A, B, C and D is $\frac{43}{240}$ letters per minute.

Hence, in a day comprising of 8 hours = 8x60 minutes, the total number of letters typed by all of them is

$$= \frac{43}{240} \times 4 \times 8 \times 60 = 344.$$

Example 9: *A man travelled 12 miles at 4 m.p.h and again 10 mile at 5 m.p.h. What was the average speed?*

Solution*:* Here, the speeds are shown in miles per hour; we are also given the miles travelled. Hence, H.M. of the speed 4 and 5 weighted by miles 12 and 10 will be the appropriate average.

x	*F*	*(f/x)*
4	12	3
5	12	2
Total	22	5

Average speed = $\frac{N}{\sum\left(\frac{f_i}{x_i}\right)} = \frac{22}{5} =$ 4.40 m.p.h.

2.5.3.1. Merits and Demerits of H.M.

Merits:

1. . H.M. is advantageous when it is desired to give greater weight to smaller observations and less weights to the larger.
2. It is used in averages involving time, rate and price.

Demerits:

1. H.M has an abstract character, and its meaning is not generally understood by common people. As such, it has only a limited use.
2. Like G.M., it cannot be calculated, if any of the observations is zero.
3. It may not be an actual value of the variable.

2.5.4. Relations between A.M., G.M. and H.M.

The A.M., the G.M. and the H.M. of a series of *n* number of observations are connected by the relation:

$$\text{A.M.} \geq \text{G.M} \geq \text{H.M.} \qquad \text{(i)}$$

The sign of equality holds if all the *n* observations are equal.

Proof: We shall establish the result (i) for two numbers only, although the result holds in general for *n* observations.

Let *a* and *b* be two real positive numbers i.e., $a > 0$, $b > 0$.

Then, $A.M = \frac{a+b}{2}$; $G.M. = \sqrt{ab}$; $H.M = \frac{2}{\frac{1}{a}+\frac{1}{b}} = \frac{2ab}{a+b}$ $(*)$

We have $A.M. - G.M. = \frac{a+b}{2} - \sqrt{ab} = \frac{a+b-2\sqrt{ab}}{2} = \frac{1}{2}\left(\sqrt{a} - \sqrt{b}\right)^2$

Since the square of a real quantity is always non-negative, we have $\left(\sqrt{a} - \sqrt{b}\right)^2 \geq 0$.

$$A.M. - G.M \geq 0 \qquad \Rightarrow \qquad A.M. \geq G.M. \qquad (ii)$$

The sign of equality holds if and if only $\left(\sqrt{a} - \sqrt{b}\right)^2 = 0 \;\Rightarrow \left(\sqrt{a} - \sqrt{b}\right) = 0 \;\Rightarrow \sqrt{a} = \sqrt{b}$

$$\Rightarrow a = b$$

i.e., if and only if the two numbers are equal.

Again we have $G.M - H.M = \sqrt{ab} - \frac{2ab}{a+b} = \sqrt{ab}\left(1 - \frac{2\sqrt{ab}}{a+b}\right) = \sqrt{ab}\left(\frac{a+b-2\sqrt{ab}}{a+b}\right)$

$$= \frac{\sqrt{ab}\left(\sqrt{a} - \sqrt{b}\right)^2}{a+b} \geq 0$$

Since $a > 0, b > 0$ and square of a real quantity is always positive.

$$\therefore G.M. - H.M \geq 0 \quad \Rightarrow \quad G.M. \geq H.M. \quad (iii)$$

The sign of equality holds if and only if $\sqrt{a} = \sqrt{b} \qquad \Rightarrow a = b$

Thus, combining (ii) and (iii), we get

$$\text{A. M.} \geq \text{G. M} \geq \text{H. M.}$$

2.5.4.1. If *a* and *b* are two positive values of a variable, their G.M. will be equal to the geometric mean of their A.M. and H.Ms.

Proof: Let *a* and *b* are two positive numbers of a variable, then from (*) above, we get

$$A.M. \times H.M. = \frac{a+b}{2} \times \frac{2ab}{a+b} = ab = (G.M.)^2$$

$$G.M. = \sqrt{ab} = \sqrt{A.M \times H.M}$$

Example 10: *The A.M of two observations is 25 and their G.M is 15. Find (i) their harmonic mean and (ii) the two observations.*

Solution: For two observations $A.M. \times H.M. = (G.M.)^2$

Substituting the values of A.M. and G.M., $25 \times H.M. = (15)^2$,

So that $H.M. = \frac{(15)^2}{25} = \frac{225}{25} = 9$

Again if *x* and *y* be the two observations $A.M. = \frac{x+y}{2} = 25 \qquad and \qquad G.M. = \sqrt{xy} = 15$

i.e., $x + y = 50$ and $xy = 15$. Substituting the value of *y* from the first equation into the second, we have

$$x(50 - x) = 225 \qquad or, \qquad 50x - x^2 = 225$$

$$or, x^2 - 50x + 225 = 0 \quad or, \quad (x-45)(x-5) = 0$$

$$\therefore x = 45, 5. \quad \text{Hence } y = 5, 45.$$

In any case, the two observations are 45 and 5.

2.6. Median

If a set of observations arranged in order of magnitudes (ascending or descending), then the middle most or central value gives the median. Median divides the group into two equal parts, one part comprising all values greater than and the other all values lesser than the median. It is not affected by the extremely large or small values. Thus, median is an average of position. Certainly, it is the real measure of central tendency.

2.6.1. Calculation of Median:

1. Median of Individual Series (Ungrouped Data)

i. Arrange the given data in ascending or descending order. For total number of observations n , then

$$Median = value\ of\ \left(\frac{n+1}{2}\right)^{th}\ observation \quad \rightarrow \{if\ n\ is\ odd\}$$

$$Median = A.M.of\ values\ of\ \left(\frac{n}{2}\right)^{th}\ and\ \left(\frac{n}{2}+1\right)^{th} \quad \rightarrow \{if\ n\ is\ even\}$$

Example 11: *Find the median of the following marks obtained by 7 students?*

4 12 7 9 14 17 16

Solution: Arrange the data in ascending order: 4, 7, 9, 12, 14, 16, 17 here, $n = 7 = an\ odd\ number$

$$\therefore Median = value\ of\ \left(\frac{n+1}{2}\right)^{th}\ observation = value\ of\ \left(\frac{7+1}{2}\right)^{th} = value\ of\ 4th\ item$$

$$\therefore Median = 12.$$

$$\therefore Median\ is\ 12\ marks.$$

Example 12: *Find the median of the following data.* 4 12 7 9 14 17 16 21

Solution: Arrangement: 4, 7, 9, 12, 14, 16, 17, 21 here, $n = 8 = an\ even\ number$

$$Median = A.M.of\ values\ of\ \left(\frac{n}{2}\right)^{th}\ and\ \left(\frac{n}{2}+1\right)^{th}$$

$$= A.M.of\ values\ of\ \left(\frac{8}{2}\right)^{th}\ and\ \left(\frac{8}{2}+1\right)^{th}\ item$$

$$= A.M.of\ values\ of\ 4^{th}\ and\ 5^{th}\ itmes$$

$$= A.M.of\ values\ of\ 12\ and\ 14\ marks$$

$$= \frac{12+14}{2} = 13\ marks$$

Alternative way:

$$Median = \ value\ of\ \left(\frac{n+1}{2}\right)^{th}\ observation = value\ of\ \left(\frac{8+1}{2}\right)^{th} = value\ of\ 4.5th\ item$$

$$= 1/2(value\ of\ 4^{th}\ item + \ value\ of\ 5^{th}\ itmes = \frac{1}{2}(12+14) = 13\ marks.$$

ii. Median for Discrete Series or Simple Frequency Distribution

The cumulative frequency (less than type) corresponding to each distinct value of the variable is calculated. If the total frequency be N, the value of variable corresponding to cumulative frequency (N+1)/2 gives the median.

Example 13: *Find the median from the following simple frequency distribution.*

x:	1	2	3	4	5	6
f:	7	12	17	19	21	24

Solution: Calculations for Median

x	*F*	*Cum. frequency*
1	7	7
2	12	19
3	17	36
4	19	55
5	21	76
6	24	100 (=N)

Now, Median = value of (N+1)/2th item = (100+1)/2th item = 50.5th item.

From 3rd column, it is found 50.5 is greater than cumulative frequency 36, but less than 55 corresponding to *x* =4. All the 19 items (from 37 to 55) have the same variate 4 and 50.5th item is also one of these 19 items.

$$\therefore Median = 4.$$

iii. Median for Continuous Series or Grouped Frequency Distribution

The formula to determine median in case of continuous data, one use N/2 and not (N+1)/2 to determine the rank of median. After locating median, its magnitude is measured by applying the formula of interpolation given below:

$$Median = l_1 + \frac{\left({}^{N}/_{2} - C\right)}{f} \times h$$

Where, l_1 = lower limit of the class in which median lies

f = Frequency of the median class

N = Total frequency i.e., $\sum f$

C = cumulative frequency of the class preceding the median class

h = Magnitude or width of the median class.

Note: While calculating the median of a series, it must be put in the 'exclusive class-interval' form. If the original series is in inclusive type, first covert it into the exclusive type and then find its median.

Example 14: *The following is the table which gives you the distribution of marks secured by some students in an examination: Find the median marks.*

Marks	0-20	21-30	31-40	41-50	51-60	61-70	71-80
No. of Students	42	38	120	84	48	36	31

Solution: Since the series is given in inclusive form, will convert it into exclusive form.

Class boundaries	*Frequency*	*Cum. frequency*
0-20.5	42	42
20.5-30.5	38	80
30.5-40.5	120	200
40.5-50.5	84	284
50.5-60.5	48	332
60.5-70.5	36	368
70.5-80.5	31	399 (= N)

Median = value of (N/2)th item = 339/2 = 199.5th item

Median lies between 30.5 and 40.5

Now, l_1 = 30.5, N = 399, C = 80, h = 10 and f = 120

$$\therefore Median = l_1 + \frac{\left({}^{N}/_{2} - C\right)}{f} \times h$$

$$= 30.5 + \frac{199.5 - 80}{120} \times 10$$

$$= 30.5 + \frac{119.5}{12}$$

$$= 30.5 + 9.96 = 40.46 \text{ Marks.}$$

Example 15: *Find the missing frequency from the following distribution of daily sales of shops, given that the median of shops is Rs. 2,400.*

Sales in hundred Rs	0-10	10-20	20-30	30-40	40-50
No. of shops	5	25	-	18	7

Solution: Let the missing frequency be 'a'

Since, median sales is Rs.2,400 (24 hundred), 20-30 is the median class. Using median formula, we get

Sales in hundred Rs	*No. of shops (f)*	*Cumulative frequency*
0-10	5	5
10-20	25	30
20-30	a	30+ a
30-40	18	48+ a
40-50	7	N = 55+ a

Here, Median = 24, N = 55 + a, $l_1 = 20, C = 30, f = a \text{ and } h = 10$

$$\therefore Median = l_1 + \frac{\left({}^{N}/_{2} - C\right)}{f} \times h$$

$$\Rightarrow 24 = 20 + \frac{\left(\frac{55+a}{2} - 30\right)}{a} \times 10$$

$$\Rightarrow 24 = 20 + \frac{55 + a - 60}{2a} \times 10$$

$$\Rightarrow 4a = 5a - 25 \Rightarrow a = 25$$

Hence the missing frequency is 25.

Example 16: *Find the median graphically from the following frequency distribution.*

Weekly wages in Rs	0-20	*20-40*	*40-60*	*60-80*	*80-100*
No. of workers	*40*	*51*	*64*	*38*	*7*

Solution: In order to know the median, we must construct an ogive of 'less than type'. For this purpose, we have to construct a cumulative frequency distribution.

Class boundary	0	20	40	60	80	100
Cum. frequency	0	40	91	155	193	200 = N

The cumulative frequencies are plotted on a graph paper against the class boundaries and the ogive is drawn (fig 2.1). From the point N/2 = 200/2 = 100 on the vertical axis, a horizontal line is drawn meeting the ogive. From the point of intersection, a perpendicular is drawn on the horizontal axis. The point showing the foot of the perpendicular is now read from scale, and is found to be approximately 43. Thus, median = 43.

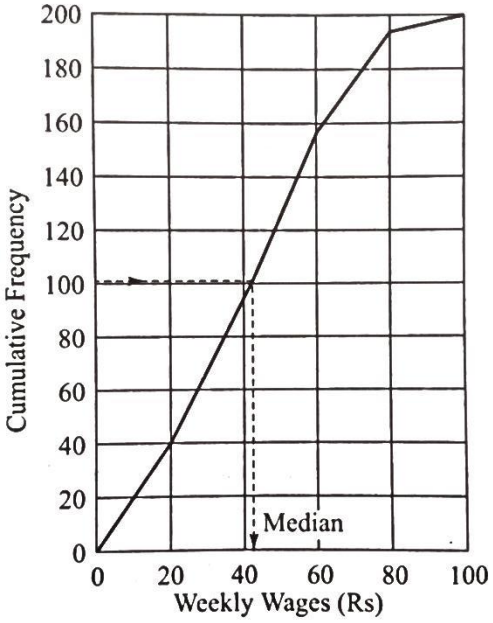


Fig 2.1. Ogive for wage distribution

2.6.2. Merits and Demerits of Median

Merits:-

1. It is rigidly defined.
2. It is easy to calculate and understand.

3. It is not affected by the extreme values like A.M.
4. It can be found by mere inspection.
5. It can be used for qualitative variables.
6. It can be obtained graphically.

Demerits:-

1. Arrangement of data is necessary while calculating median.
2. It cannot be used to calculate the combined medians of two or more groups.
3. It is affected more by sampling fluctuations than the A.M.

2.6.3. PARTITION VALUES- QUARTILES, DECILES AND PERCENTILES

The partition values divide the observation into different equal parts. Just as median, which divides the series into two equal parts, there are some other measures which divide the series into like four, ten or hundred equal parts. The median divides the series into two equal parts, quartiles divides into four equal parts and the percentile into hundred equal parts.

Quartiles: Quartiles divides the series into four equal parts. The number of observations between each quartile is the same. Obviously, there are three quartiles

i. *First quartile (or Lower quartile): Q_1*
ii. *Second quartile (or Middle quartile or median): Q_2*
iii. *Third quartile (or Upper quartile): Q_3*

$$Q_1 < Q_2 < Q_3;\ Q_2 = \text{Median}$$

Deciles: Deciles are such values which divides the total number of observations into ten equal parts. There are nine deciles $D_1, D_2, D_3, \ldots\ldots, D_9$ called the first decile, second decile etc.

$$D_1 < D_2 \ldots\ldots < D_9;\ \ D_5 = Q_2 = \text{Median}$$

Percentiles: Percentiles are such values which divides the total number of observations into hundred equal parts. There are 99 percentiles $P_1, P_2, \ldots\ldots, P_{99}$, called the first percentile, second percentile etc.

$$P_{10} = D_1, \quad P_{20} = D_2, \ldots\ldots, P_{90} = D_9$$

$$P_{25} = Q_1, \quad P_{50} = D_5 = Q_2 = \text{Median}, P_{75} = Q_3, \quad P_1 < P_2 < \cdots < P_{99}.$$

Calculation of partition values:

(a) For Simple Series: The data are to be arranged in ascending order of magnitude.

1st Quartile, $Q_1 =$ value of $\frac{n+1}{4}$th item; 3rd Quartile, $Q_3 =$ value of $\frac{3(n+1)}{4}$th item

1st Decile, D_1 = value of $\frac{n+1}{10}$ th item; 8th Decile, D_8 = value of $\frac{8(n+1)}{10}$ th item

K − th Decile, D_k = value of $\frac{k(n+1)}{10}$ th item (for k = 1, 2, 3, … …9)

1st percentile, P_1 = value of $\frac{n+1}{100}$ th item,

K − th Percentile, P_k = value of $\frac{k(n+1)}{100}$ th item (for k = 1, 2, 3, … …99)

Example 17: *Find the values of Q_1, Q_3, D_4 and P_{60} from the following weights.*

19, 47, 27, 24, 39, 26, 34, 42, 45, 44, 51, 67, 50, 59, 57, 60, 59, 62, 56, 57.

Solution: Arrangement.

Serial number	*Weight (kg.)*	*Serial number*	*Weight (kg.)*
1	19	11	50
2	24	12	51
3	26	13	56
4	27	14	57
5	34	15	57
6	39	16	59
7	42	17	59
8	44	18	60
9	45	19	62
10	47	20	67

Here $n = 10$

Q_1 = value of $\frac{n+1}{4}$ th item = value of $\frac{20+1}{4}$ th item = value of 5.25th item

$$= \text{value of 5th item} + \frac{1}{4}(\text{value of 6th item} - \text{value of 5th item})$$

$$= 34 + \frac{1}{4}(39 - 34) = 34 + 1.25 = 35.25 \text{ kg.}$$

Q_3 = value of $\frac{3(n+1)}{4}$th item = value of $\frac{3(20+1)}{4}$th item = value of 15.75th item

= value of 15th item + $\frac{3}{4}$(value of 16th item − value of 15th item)

= $57 + \frac{3}{4}(59 - 57) = 57 + 1.50 = 58.50$ kg.

D_4 = value of $\frac{4(n+1)}{10}$th item = value of $\frac{4(20+1)}{10}$th item = value of 8.4th item

= value of 8th item + $\frac{4}{10}$(value of 9th item − value of 8th item)

= $44 + \frac{4}{10}(45 - 44) = 44 + 0.4 + 44.4$ kg.

P_{60} = value of $\frac{60(n+1)}{100}$th item = value of $\frac{60(20+1)}{100}$th item = value of 12.6th item

= value of 12th item + $\frac{6}{10}$(value of 13th item − value of 12th item)

= $51 + \frac{6}{10}(56 - 51) = 51 + 3 = 54$ kg.

(b) For Simple Frequency Distribution: The less than cumulative frequency corresponding to each distinct value of the variable is calculated. If the total frequency be N.

Q_1 = value of $\frac{N+1}{4}$ th item

Q_3 = value of $\frac{3(N+1)}{4}$ th item

D_k = value of $\frac{k(N+1)}{10}$ th item

P_k = value of $\frac{k(N+1)}{100}$ th item

Example 18: *Find the values of* $Q_1, Q_3,\ D_4$ and P_{60} *from the following frequency distribution.*

Weight (kg.)	40	42	45	50	51	54	56	59	60	62	64
Frequency	2	6	8	10	6	14	12	8	14	12	6

Solution: In order to derive the values of quartiles, deciles and percentiles we have to construct the cumulative frequency of less than type.

Calculations for Quartiles, Deciles and Percentiles

Weights (Kg.)	*Frequency*	*Cumulative Frequency*
40	2	2
42	6	8
45	8	16
50	10	26
51	6	32
54	14	46
56	12	58
59	8	66
60	14	80
62	12	92
64	6	98 = N

Q_1 = value of $\frac{N+1}{4}$ th item = value of $\frac{98+1}{4}$ th item = value of 54.75th item = 50 kg.

Q_3 = value of $\frac{3(N+1)}{4}$ th item = value of $\frac{3(98+1)}{4}$ th item = value of 74.25th item = 60 kg.

D_4 = value of $\frac{4(N+1)}{10}$ th item = value of $\frac{4(98+1)}{10}$ th item = value of 39.6th item = 54 kg.

P_{60} = value of $\frac{60(N+1)}{100}$ th item = value of $\frac{60(98+1)}{100}$ th item = value of 59.4th item = 59 kg.

(c) For Continuous Series: Like Median, the values of quartiles, deciles and percentiles lie in various class-intervals and the actual values are to be calculated by applying interpolation formulae.

Q_1 = value of $\frac{N}{4}$ th item , $\therefore \quad Q_1 = l_1 + \frac{(N/4 - C)}{f} \times h$

Q_3 = value of $\frac{3N}{4}$ th item , $\therefore \quad Q_3 = l_1 + \frac{(3N/4 - C)}{f} \times h$

D_k = value of $\frac{kN}{10}$ th item , $\therefore \quad D_k = l_1 + \frac{(kN/10 - C)}{f} \times h$ (for k = 1, 2, 3,……., 9)

P_k = value of $\frac{kN}{100}$ th item , $\therefore \quad P_k = l_1 + \frac{(kN/100 - C)}{f} \times h$ (for k = 1, 2, 3,……., 99)

***Example 19**: Find the values of 1st and 3rd quartiles, 6th decile and 70th percentile from the following frequency distribution.*

Marks	*No. of students*	*Marks*	*No. of students*
Less than10	*5*	*Less than 50*	*60*
" 20	*13*	*" 60*	*80*
" 30	*20*	*" 70*	*90*
" 40	*32*	*" 80*	*100*

***Solution**:* The cumulative frequency is given; we need to convert it into the grouped frequency distribution. We also convert the above distribution into exclusive type classes with class boundaries below 9.5, and so on.

Class	*Class boundary*	*Frequency*	*Cumulative frequency*
Less than 10	Below 9.5	5	5
10-19	9.5-19.5	13-5 = 8	13
20-29	19.5-29.5	20-13 = 7	20
30-39	29.5-39.5	32-20 = 12	32
40-49	39.5-49.5	60-32 = 28	60
50-59	49.5-59.5	80-60 = 20	80
60-69	59.5-69.5	90-80 = 10	90
70-79	69.5-79.5	100-90 = 10	100 = N

For $Q_1 =$ value of $\frac{N}{4}$th item $= \frac{100}{4} =$ value of 25th item, here Q_1 lies between 29.5-39.5 class interval.

$$Q_1 = l_1 + \frac{\left({}^{N}/_{4} - C\right)}{f} \times h$$

$Here\ N = 100, \quad l_1 = 29.5, \quad f = 12, \quad C = 20 \quad and\ h = 10$

$$\therefore \quad Q_1 = 29.5 + \frac{(25-20)}{12} \times 10 = 29.5 + \frac{5}{12} \times 10 = 29.5 + 4.17 = 33.67$$

$For\ Q_3 =$ value of $\frac{3N}{4}$ th item $= \frac{3\times100}{4} =$ value of 75th item, here Q_3 lies between 49.5-59.5 class interval.

$$Q_3 = l_1 + \frac{\left(3N/_4 - C\right)}{f} \times h$$

$$Here\ l_1 = 49.5, \quad f = 20, \quad C = 60\ \ and \quad h = 10$$

$$\therefore\ Q_3 = 49.5 + \frac{(75-60)}{20} \times 10 = 49.5 + \frac{15}{20} \times 10 = 49.5 + 7.5 = 57.0.$$

$For\ D_6 =$ value of $\frac{6N}{10}$ th item $= \frac{6\times100}{10} =$ value of 60th item, here D_6 lies between 49.5-59.5 class interval.

$$D_6 = l_1 + \frac{\left(6N/_{10} - C\right)}{f} \times h$$

$$Here\ l_1 = 49.5, \quad f = 20, \quad C = 60\ \ and \quad h = 10$$

$$\therefore D_6 = 49.5 + \frac{(60-60)}{20} \times 10 = 49.5.$$

$For\ P_{70} =$ value of $\frac{70N}{100}$ th item $= \frac{70\times100}{100} =$ value of 70th item, here D_6 lies between 49.5-59.5 class interval.

$$P_{70} = l_1 + \frac{\left(70N/_{100} - C\right)}{f} \times h$$

$$Here\ l_1 = 49.5, \quad f = 20, \quad C = 60\ \ and \quad h = 10$$

$$\therefore\ P_{70} = 49.5 + \frac{(70-60)}{20} \times 10 = 49.5 + 5 = 54.5.$$

Example 20: *Calculate the 3rd decile and quartiles graphically for the following data.*

Height (inches)	57.5-60.5	60.0-62.5	62.5-65.0	65.0-67.5	67.5-70.0	70.0-72.5	72.5-75.0
No. of people	6	26	190	281	412	127	38

Solution: Firstly we calculate the cumulative frequencies.

Class boundaries	*Cumulative frequency (less than)*
57.5	0
60.0	6
62.5	32
65.0	222

67.5	503
70.0	915
72.5	1042
75.0	1080=N

We draw the ogive from the data given in Q.19. We know, $Q_1 = \frac{N}{4} = \frac{1080}{4} = 270, Q_2 = \frac{N}{2} = \frac{1080}{2} = 540,$

$$Q_3 = \frac{3N}{4} = 3 \times \frac{1080}{4} = 810 \; and \; D_3 = \frac{3N}{10} = 3 \times \frac{1080}{10} = 324.$$

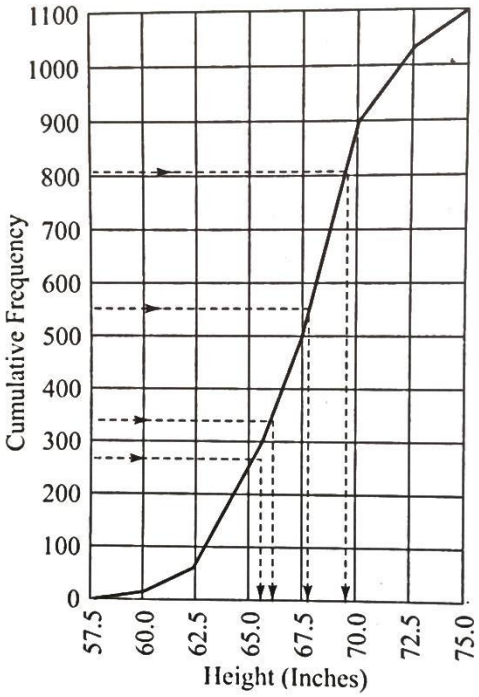


Fig 2.2: Ogive and location of partition values

Thus, we draw the horizontal lines from cumulative frequencies 270, 324, 540 and 810 on the vertical scale meeting the ogive. The abscissae of the points of intersection are read approximately from the horizontal scale, giving the values of Q_1, Q_2, Q_3, D_3 respectively. Thus, we find approximately $Q_1 = 65.43$ inches, $Q_2 = 67.7$ inches, $Q_3 = 69.3$ inches and $D_3 = 65.9$ inches.

2.7. MODE

It is the size of that item which possesses the maximum frequency. According to Kenney and Keeping, the value of the variable occurs most frequently in a distribution is called the mode. It is the most common value. It is the point of maximum density.

2.7.1 Calculations of Mode:

(a) Ungrouped data:

Individual series: The mode of this type of series can be obtained by mere inspection. The number which occurs most often is the mode.

Example 21: *Find the mode from the following given data.*

x	1	2	3	4	5	6	7	8	9
f	3	1	18	25	40	30	22	10	6

Solution: It is clear that the observation with the value 5 has occurred for maximum numbers of timew that is 40 times. Hence, the mode is 5.

(b) Grouped data: For a continuous frequency distribution, the class corresponding to the maximum frequency is called the modal class and the value of mode is obtained by the interpolation formula:

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where l is the lower class of the modal class

f_1 is the frequency of the modal class or the maximum frequency

f_0 is the frequency of the class preceding the modal class

f_2 is the frequency of the class succeeding the modal class

h is the magnitude of the modal class.

Example 22: *Find the mode of the following frequency distribution.*

Class intervals	50-59	60-69	70-79	80-89	90-99	100-109
Frequency	5	25	65	115	145	151

Solution: In the above table the class interval is of inclusive type, we have to convert it into an exclusive type. Again, the cumulative frequencies are given, we have to find out the grouped frequencies.

Class boundaries	49.5-59.5	59.5-69.5	69.5-79.5	79.5-89.5	89.5-99.5	99.5-109.5
Frequencies	5	20	40	50	30	6

Since the highest frequency is 50, so modal class is (79.5-89.5).

Here $l = 79.5, f_1 = 50, f_0 = 40, f_2 = 30$ and $h = 10$

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h = 79.5 + \frac{50 - 40}{100 - 40 - 30} \times 10 = 79.5 + \frac{10}{30} \times 10 = 79.5 + 3.33 = 82.83$$

Example 23: *The monthly profits of 100 shops are distributed as follows:*

Profit per shop	0-100	100-200	200-300	300-400	400-500	500-600
No. of shops	12	18	27	20	17	6

Draw the histogram to the data and find the modal value. Check this value by direct valuation.

Solution: First we will draw the histogram to the data given above. Here the modal class is 200 – 300, so it gives the highest rectangle. Now, in the inside of this highest rectangle draw the line diagonally starting from the upper corner of the bar to the upper corner of the adjacent bar. Now, draw a perpendicular from the point of intersection of the diagonal lines to the x-axis. The point is thus, read off gives the modal value and it is approximately equal to 256 rupees.

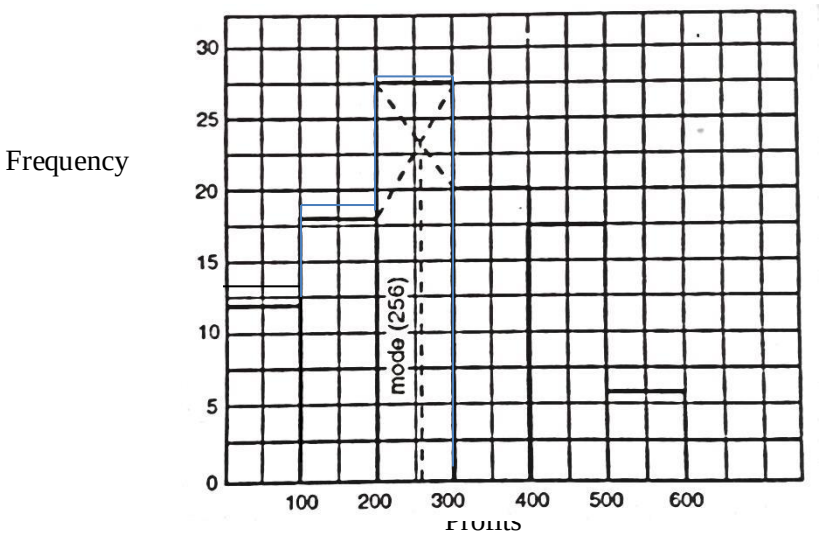


Fig 2.3: Histogram showing the distribution of profits

From the direct calculation we found modal class is 200-300.

Here $l = 200, f_1 = 27, f_0 = 18, f_2 = 20$ and $h = 100$

$$\therefore Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h = 200 + \frac{27 - 18}{54 - 18 - 20} \times 100 = 200 + \frac{9}{16} \times 100$$

$$= 200 + 56.25 = Rs.256.25$$

2.5.2 Merits and Demerits of Mode

Merits:-

1. It is simple to calculate.
2. In individual or discrete distribution it can be located by mere inspection.
3. It is not affected by sampling fluctuations. Mode can be determined graphically.

Demerits:-

1. It is ill defined.
2. It is not based on all observations.
3. It is not capable of further algebraic treatment.
4. It is not a good representative of the data.
5. Sometimes there are more than one values of mode.

2.8. Empirical Relation between Mean (M), Median (M_d) and Mode (M_o)

A distribution in which the values of mean, median and mode coincide, is known symmetrical and if above values are not equal then the distribution is said asymmetrical or skewed. In a moderately skewed distribution, there is a relation between mean, median and mode which is as follows:

$$Mean - Mode = 3(Mean - Median)$$

If any two values are available we can find the third.

Example 24: *If A.M = 35.4 and Mode = 32.1. Find Median.*

Solution: We know

$$Mean - Mode = 3(Mean - Median)$$

$$\Rightarrow\ 3\,Median = 2\,Mean + Mode$$

$$\Rightarrow Median = \frac{2 \times 35.4 + 32.1}{3} = \frac{102.9}{3} = 34.3.$$

2.9. Best Measure of Average

From the above discussion of measure of central tendency, we find that each measure has its own advantages and limitations and consequently has its own field of importance and utility. For example, arithmetic mean cannot be used the distribution has extreme values or open end classes. For open end classes, median and mode can be used as an average. Median is very useful in qualitative variables. Geometric mean is usually used with rates and ratios and harmonic mean is used when time factor of variable is given. Thus, we cannot use averages indiscriminately. The selection of an average depends upon certain conditions such as the purpose of the enquiry, the nature and availability of data, types of variables involved, the method of classification adopted etc.

However, since Arithmetic Mean satisfies almost all the properties of a good average. It may be considered as the best available so far.

Exercises

1. Calculate arithmetic mean and median from the following series: [C.S. (Foundation), Dec 2000]

Income (Rs)	0-5	5-10	10-15	15-20	20-25	25-30
Frequency	5	7	10	8	6	4

Ans: A.M. = 14.37; Median = 14.

2. Find the 45^{th} and 57^{th} percentiles for the following data on marks obtained by 100 students. [C.A. (Foundation), May 1996]

Marks	20-25	25-30	30-35	35-40	40-45	45-50
No. of Students	10	20	20	15	15	20

Ans: $P_{45} = 33.75$; $P_{57} = 37.33$

3. With the help of the following data, find graphically using ogive, the values of Q_1, Q_3, Median, P_{40} and D_6. [D.U., B.Com. (Hons), 2008]

Class interval	10-14	15-19	20-24	25-29	30-34	35-39	Total
Frequencies	5	10	15	20	10	5	65

Ans: $Q_1 = 19.92$; $Q_3 = 29.19$; Median = 25.13; $P_{40} = 23.17$; $D_6 = 26.75$.

4. An incomplete distribution is given below: [Himachal Pradesh Univ. B.Com, 1999; Kerala Univ. B.Com., 1999]

Variable	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	10	20	?	40	?	25	15

(i) The median is 35. Find out missing frequency (given that total frequency = 170).

(ii) Calculate the arithmetic mean of the completed table.

Ans: (i) 35, 25 (ii) 35.88

5. The average marks obtained in an examination by two groups of students was found to be 75 and 85 respectively. Determine the ratio of students in the two groups, if the average mark for all students was 8.

[B.U., B.A (Econ) 1969]

Ans: 1 : 1.

6. In moderately skewed distribution, Mean = 24.6 and the mode = 24. Find the median. [C.A., Nov. 1967]

7. The G.M., H.M. and A.M. of three observations are 3.63, 3.27 and 4 respectively. Find the observations.

Ans: 2, 4, 6.

UNIT III MEASURES OF DISPERSION

3.1. Objectives: - In the last chapter we learn how a single value can represent the whole distribution. In this section we shall come to know how the observations are scattered around that central value.

3.2. Introduction: We may find the observations scattered around the central value where some are less and some are higher than the average. Dispersion means how the observations scattered around the central value.

When the values of observations are closed to the value of mean we say the dispersion is less and when the observations are far from mean we say dispersion is high and when all the observations are same as mean then, the dispersion is zero. For, example,

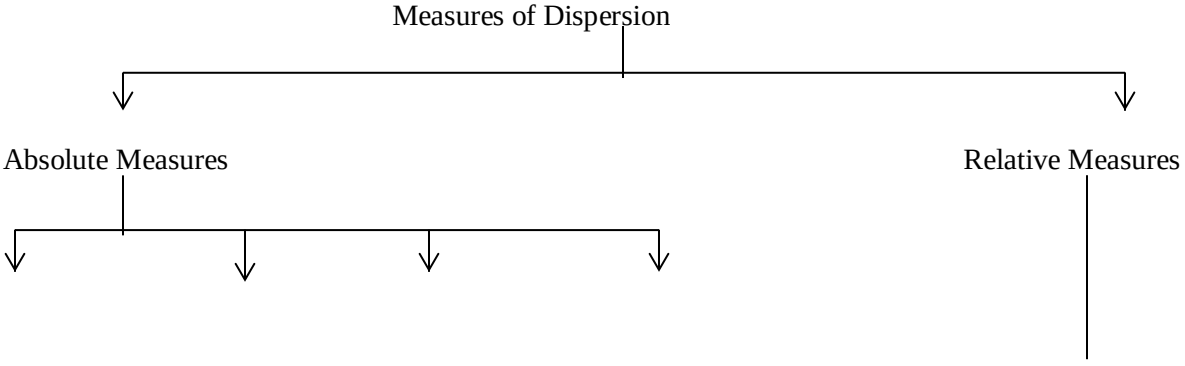
1st set of observations	x_i	4	4	4	4	Mean = 4
2nd set of observations	x_i	6	3	2	5	Mean = 4
3rd set of observations	x_i	8	1	1	6	Mean = 4

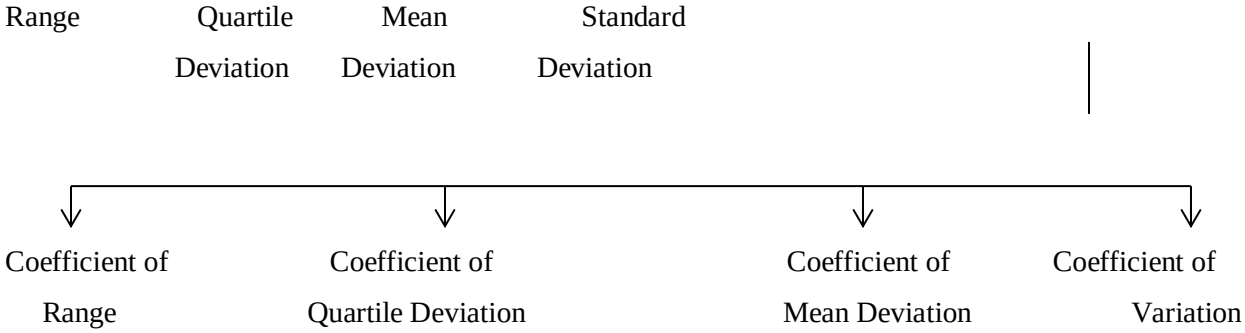
In the 1st set of observations the mean is 4 and each observation is also 4. In this case we say the distribution has zero dispersion. In the 2nd set of observations the observations are more or less close to the mean value whereas in the 3rd set of observations the observations are far from the value of mean. Thus, we see that even if the mean is same for all three set of observations, for 2nd and 3rd third set of observations the values of each observation is different from that of mean. Thus, the observations are dispersed around the mean and we say for 2nd example the dispersion is less while for 3rd example the dispersion is higher. Thus, sometimes mean is not enough to explain the variation available in the data. Hence, the analysis of dispersion or variation comes into consideration.

3.3. Properties of a Measure of Variation:

i. It should be based on all the observations.
ii. It should be simple to understand and easy to compute.
iii. It should be rigidly defined.
iv. It should not be affected by the extreme values.
v. It should have sampling stability.

3.4. Methods of Studying Dispersion: There are several absolute and relative measures of dispersion. Lesser the values of these measures, lesser are the dispersion. Thus, small value is better. If the dispersion of a set of observations, say, heights of students of a class in absolute figures, then dispersion will also be in the same unit, i.e., heights. This is absolute dispersion. If, dispersion is calculated as a ratio (percentage) of the average, then it is relative dispersion. Relative dispersions are free of units of measurement and can be used to compare two or more distribution. These measures are listed below.





3.5. Range

For a set of observations the difference between the maximum and the minimum value of observations is known as range.

Thus, ***Range = Maximum Value – Minimum Value.***

Example1: *Find the range of the following marks of 8 students.*

85, 65, 97, 12, 35, 45, 9, 56

Solution: Arranging marks in an ascending order, we have

9, 12, 35, 45, 56, 65, 85, 97

$\therefore Range = Maximium\ value - Minimum\ value$

$= 97 - 9 = 88\ marks.$

Example 2: *Find the range of the following data.* *[D.S.W., Nov '73]*

Height (inches)	60-62	63-65	66-68	69-71	72-74
No. of Students	8	27	42	18	5

Solution: *In case grouped frequency distribution, the frequency of the class does not matter. Range will be calculated from class boundaries. Since class interval is given, we will first change it class boundaries.*

Class boundaries	59.5-62.5	62.5-65.5	65.5-68.5	68.5-71.5	71.5-74.5
Frequency	8	27	42	18	5

$$\therefore Range = Upper\ limit\ of\ highest\ class\ boundary - Lower\ limit\ of\ lowest\ boundary$$

$$= 74.5 - 59.5 = 15\ inches.$$

3.5.1. Coefficient of Range

The formula for coefficient of range is given as

Coefficient of Range = $\frac{Range}{Maximum\ value+Minimum\ value} = \frac{L-S}{L+S}$ where L is the largest and S is the smallest value.

Example 3: *Calculate the Coefficient of Range using the data given in first example.*

Solution: The Maximum and Minimum values in example one are 97 and 9.

$$\therefore Coefficient\ of\ Range = \frac{L-S}{L+S} = \frac{97-9}{97+9} = 0.83$$

3.5.2. Merits and Demerits of Range:

Merits:

1. Range is very easy to calculate and understand.

Demerits:

1. Range is not based on all observations. Its value can be influenced by extreme values. It cannot be calculated on open end distribution.

3.6. Quartile Deviation

Quartile deviation is defined as the half of the difference between the upper Quartile and the lower quartile.

$$\therefore Quartile\ Deviation = \frac{Q_3 - Q_1}{2}$$

The difference $Q_3 - Q_1$ is known as the interquartile range and $\frac{Q_3-Q_1}{2}$ is known as Semi interquartile range.

The quartile deviation ignores the largest 25% of the smallest observation and 25% of the largest observations. It is therefore unaffected by the extreme values. The quartile deviation can easily be calculated using simple interpolation.

3.6.1. Coefficient of Quartile Deviation: Quartile deviation is the absolute measure of dispersion. When semi quartile range divided by the average of two quartiles, we get the Coefficient of Quartile Deviation.

Symbolically, *Coefficient of Quartile Deviation* $= \frac{\frac{Q_3-Q_1}{2}}{\frac{Q_3+Q_1}{2}} = \frac{Q_3-Q_1}{Q_3+Q_1}$.

Example 4: *Find the quartile deviation and coefficient of quartile deviation of the following observations.*

(Scores): 12, 14, 15, 17, 19, 21, 25, 28, 30, 31, 33

Solution: Here *n* = 11, the observations are arranged in order.

$$Q_1 = value\ of\ \frac{n+1}{4} th\ item = \frac{11+1}{4} = value\ of\ 3rd\ item = 15\ scores$$

$$Q_3 = value\ of\ \frac{3(n+1)}{4} th\ item = \frac{3(11+1)}{4} = value\ of\ 9th\ item = 30\ scores$$

Quartile Deviation (Q.D) $= \frac{Q_3 - Q_1}{2} = \frac{30-15}{2} = 7.5\ scores$

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{30-15}{30+15} = \frac{15}{45} = 0.33$

Example 5: *Calculate the appropriate measure of dispersion from the following data.* [I.C.W.A., Jan '64]

Wages in rupees per week	*Number of wage earners*
Less than 35	14
35-37	62
38-40	99
41-43	18
Over 43	7

Solution: Since in the frequency distribution classes are open ended, so Q.D would be the appropriate measure of dispersion.

Wages (Rs.)	*Class boundaries*	*Frequency (f)*	*Cum. frequency*
Less than 35	Less than 34.5	14	14
35-37	34.5-37.5	62	76
38-40	37.5-40.5	99	175
41-43	40.5-43.5	18	193
Over 43	Over 43.5	7	200 (= *N*)

$$Q_1 = value\ of \frac{N}{4} the\ item = value\ of\ 50th\ item$$

$$Q_1\ lies\ between\ class\ (34.5 - 37.5\ )$$

$$\therefore Q_1 = 34.5 + \frac{(34.5 - 37.5)}{62}(50 - 14) = Rs.\ 36.24$$

$$Q_3 = value\ of \frac{3N}{4} th\ item = value\ of\ 150th\ item$$

$$Q_3\ lies\ between\ class\ (37.5 - 40.5\ )$$

$$\therefore Q_3 = 37.5 + \frac{(37.5 - 40.5)}{99}(150 - 76) = Rs.39.74$$

$$\therefore Quartile\ Deviation\ = \frac{Q_3 - Q_1}{2} = \frac{39.74 - 36.24}{2} = 1.75$$

$$Again, Coefficient\ of\ Quartile\ Deviation = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{39.74 - 36.24}{39.74 + 36.24} = 0.046$$

3.6.2. Merits and Demerits of quartile Deviation

Merits

1. It is superior to range as a measure of dispersion.
2. In case of open-end distribution, the quartile deviation is an appropriate measure.
3. It I not affected by the extreme values.

Demerits

1. Quartile deviation is not based on all observation and further treatment is also not possible.
2. Its value is much effected by sampling fluctuations
3. It I not a measure of dispersion, particularly for series in which variation is considered.

3.7. Mean Deviation (or Average Deviation or Mean Absolute Deviation)

Range and Quartile deviation are based on only two values. Range on extremes and quartile deviation on quartiles only. These two measures are not based on all observations. They ignore most of it. Mean deviation and Standard deviation are two measures of dispersion based on all the observations. Thus, these two measures are far superior to Range and Quartile deviation.

Definition: Mean deviation of a set of observations is the arithmetic mean of the absolute deviations of various items from the mean or median of that set of observation.

Mean deviation about median is preferred to mean deviation about mean since the mean deviation about median is less than that of about mean. The formula for mean deviation is given as

1. For simple series

$$Mean\ Deviation = \frac{1}{n} \sum |x_i - \bar{x}|$$

2. For frequency distribution

Mean Deviation = $\frac{1}{N}\sum f_i|x_i - \bar{x}|$

3.7.1. Coefficient of Mean Deviation

Coefficient of M.D, about mean = $\frac{M.D}{Mean}$

Coefficient of M.D, about median = $\frac{M.D}{Median}$

Example 6: *Find the mean deviation about median of the following data.* [B.U. B.Com 1977]

13, 84, 68, 24, 96, 139, 84, 27

Solution:

Since there are even numbers of observations, viz. 8, the median is the average of the two middlemost observations, when arranged in order of magnitude: 13, 24, 27, 68, 84, 84, 96, 139.

∴ Median = $\frac{68+84}{2} = 76$.

X	$\|x - median\|$
13	63
84	8
68	8
24	52
96	20
139	63
84	8
27	49
	$\sum\|x - median\| = 271$

Mean Deviation about median = $\frac{1}{n}\sum|x - median| = \frac{271}{8} = 33.88$

Example 7: *Find the median and mean deviation of the following data:* [Mysore Univ. B.Com., 1998]

Size	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	7	12	18	25	16	14	8

Solution: Let x_i represent mid values of classes.

Class Boundaries	*F*	*Mid- value (x)*	*Cum. freq*	$\|x - median\|$	$f\|x - median\|$
0-10	7	5	7	30.2	211.4
10-20	12	15	19	20.2	242.4
20-30	18	25	37	10.2	183.6
30-40	25	35	62	0.2	5
40-50	16	45	78	10.2	163.2
50-60	14	55	92	20.2	282.8
60-70	8	65	100 (= N)	30.2	241.6
	$\sum f$ =100				$\sum f\|x - 35.2\|$= 1330

Median = value of N/2 th item = value of 50th item. Therefore median lies between (30 – 40) class.

$$Median = l_1 + \frac{\left(N/2 - C\right)}{f} \times h = 30 + \frac{50 - 37}{25} \times 10 = 35.2$$

$$Mean\ Deviation = \frac{\sum f|x - median|}{\sum f} = \frac{1330}{100} = 13.30$$

3.7.2. Merits and Demerits of Mean Deviation

Merits

1. Mean Deviation is based on all the observations and any change in any item would change the value of mean deviation.
2. It is simple to understand and easy to compute.
3. It is less affected by the extreme values.

Demerits

1. It ignores the algebraic signs of the deviation, and as such is not capable of further algebraic treatment.
2. It is not an accurate measure, particularly when it is calculated from mode.

3. It is less popular than standard deviation.

3.8. Standard Deviation

Definition: Standard deviation of a set of observations is defined as the positive square root of the arithmetic mean of squares of all the deviations from the mean. In short, it may be defined as Root-Mean-Square-Deviation from mean. It is denoted by the Greek letter σ (sigma).

In case of mean deviation, the absolute values of all deviations are considered which is illogical in terms of mathematical language. Thus, the standard deviation is superior to mean deviation.

If $x_1, x_2, \dots\dots\dots, x_n,$ be the set of observation and $\bar{x}$ be their mean then,

$$\text{S.D}\,(\sigma) = \sqrt{\frac{1}{n}\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\}} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$$

For frequency distribution

$$\text{S.D}\,(\sigma) = \sqrt{\frac{1}{N}\{f_1(x_1-\bar{x})^2 + f_2(x_2-\bar{x})^2 + f_3(x_3-\bar{x})^2 + \cdots + f_n(x_n-\bar{x})^2\}} = \sqrt{\frac{\sum f_i(x_i-\bar{x})^2}{N}}$$

The square of S.D is known as variance, thus it is denoted by

For simple series

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

$$= \frac{1}{n}\sum {x_i}^2 - 2\frac{1}{n}\sum x_i\bar{x} + \frac{1}{n}\sum\bar{x}^2 = \frac{\sum {x_i}^2}{n} - 2\bar{x}^2 + \bar{x}^2$$

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

For frequency distribution

$$\text{Variance}\,(\sigma^2) = \frac{1}{N}\sum f_i(x_i-\bar{x})^2$$

$$= \frac{1}{N}\sum f_i {x_i}^2 - 2\frac{1}{N}\sum f_i x_i\bar{x} + \frac{1}{N}\sum f_i\bar{x}^2 = \frac{\sum f_i {x_i}^2}{N} - 2\bar{x}^2 + \bar{x}^2$$

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

Thus, S.D. is the positive square-root of Variance.

3.8.1. Properties of Standard Deviation

1. Standard deviation is independent of change of origin.

Proof: For the *n* observations $x_1, x_2, \dots\dots, x_n$ and let $y_1, y_2, \dots\dots, y_n$ be respective quantities obtained by shifting the origin to any arbitrary constant, say *c*, so that $y_i = x_i - c$.

Then, $\sigma_x{}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

Now, $y_i = x_i - c$, so that $\sum y_i = \sum x_i - \sum c$ (taking $\sum$ to both sides)

Dividing both sides by *n*, we get

$$\frac{\sum y_i}{n} = \frac{\sum x_i}{n} - \frac{\sum c}{n} \qquad \Rightarrow \qquad \bar{y} = \bar{x} - c \qquad or, \qquad \bar{x} = \bar{y} + c$$

Now, $x_i - \bar{x} = (c + y_i) - (\bar{y} + c) = y_i - \bar{y}$

So, $\sigma_x{}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 = \sigma_y{}^2$

$\therefore \sigma_x = \sigma_y$

2. Standard deviation is independent of change of origin, but is dependent on the change of scale.

Proof: For the *n* observations $x_1, x_2, \dots\dots, x_n$ and let the origin be changed to *c* and the scale to *d*, then

$y_i = \frac{x_i - c}{d}$ $or, x_i = c + dy_i,$ which means $y_1, y_2, \dots\dots, y_n$ are the deviations of $x_1, x_2, \dots\dots, x_n$ from an arbitrary constant *c*, in units of another constant *d*.

Now, $\bar{x} = d\bar{y} + c$

Again $x_i - \bar{x} = (c + dy_i) - (d\bar{y} + c) = d(y_i - \bar{y})$

$$\sigma_x{}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum d^2(y_i - \bar{y})^2 = d^2\sigma_y{}^2$$

$$\therefore \sigma_x = d\sigma_y$$

3. If a group of n_1 observations has mean $\bar{x}_1$ and S.D. σ_1 and another group of observations n_2 has mean $\bar{x}_2$ and S.D. σ_2, then S.D (σ) of the composite group of $n_1 + n_2$ $(= N)$ observations can be obtained by the formula

$$N\sigma^2 = (n_1\sigma^2{}_1 + n_2\sigma^2{}_2) + (n_1d^2{}_1 + n_2d^2{}_2)$$

Where $d_1 = \overline{x_1} - \bar{x}$, $d_2 = \overline{x_2} - \bar{x}, and\ N\bar{x} = n_1\bar{x}_1 + n_2\bar{x}_2$.

$$\sigma = \sqrt{\frac{(n_1\sigma^2{}_1 + n_2\sigma^2{}_2) + (n_1d^2{}_1 + n_2d^2{}_2)}{N}} = \sqrt{\frac{(n_1\sigma^2{}_1 + n_2\sigma^2{}_2) + (n_1d^2{}_1 + n_2d^2{}_2)}{n_1 + n_2}}$$

The above relation can be extended to any number of groups

$$\sigma = \sqrt{\frac{\sum n_i {\sigma^2}_i + \sum n_i {d^2}_i}{\sum n_i}}$$

4. Mean and standard deviation of the first natural numbers are $\frac{n+1}{2}$ *and* $\sqrt{\frac{n^2-1}{12}}$

Proof: Let first natural numbers are 1, 2, 3, 4,……., *n*. The sum and the sum of squares of these *n* numbers are

$$\sum x = 1 + 2 + 3 + 4 + \cdots + n = \frac{n(n+1)}{2}$$

$$\sum x^2 = 1^2 + 2^2 + 3^2 + 4^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$Mean(x) = \frac{\sum x}{n} = \frac{n(n+1)}{2n} = \frac{(n+1)}{2}$$

$$\sigma^2 = \frac{\sum {x_i}^2}{n} - \bar{x}^2 = \frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4}$$

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

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

$$= \frac{(n+1)\{4n+2-3n-3\}}{12}$$

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

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

3.8.2. Coefficient of standard deviation$= \frac{\text{Standard Deviation}}{\text{Mean}}$

3.8.3. Coefficient of variation $= \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$

Smaller the value of coefficient of variation, the more consistent the variable is.

Example 8: *Find the standard deviation of the given data,*

9, 7, 5, 11, 1, 5, 7, 3 [N.B., B.Com. 1981]

Solution: Calculations for S.D

X	x^2
9	81
7	49
5	25
11	121
1	1
5	25
7	49
3	9
$\sum x = 48$	$\sum x^2 = 360$

Mean = $\frac{\sum x}{\text{n}} = \frac{48}{8} = 6$

S.D = $\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{\text{n}}\right)^2} = \sqrt{\frac{360}{8} - \left(\frac{48}{8}\right)^2} = \sqrt{45 - 36} = \sqrt{9} = 3.$

Example 9: *Find the standard deviation of the following distribution:* [C.U., B.A. (Econ) 1974]

X	0-500	500-1000	1000-1500	1500-2000	2000-3000
Frequency	90	218	86	41	15

Solution: Calculations for S.D.

X	*f*	*Mid-value (x)*	$y = \frac{x-750}{250}$	*Fy*	fy^2
0-500	90	250	-2	-180	360
500-1000	218	750	0	0	0
1000-1500	86	1250	2	172	344
1500-2000	41	1750	4	164	656

2000-3000	15	2500	7	105	735
	$N = 450$			$\sum fy = 261$	$\sum fy^2 = 2095$

$$\sigma_y{}^2 = \frac{\sum fy^2}{N} - \left(\frac{\sum fy}{N}\right)^2 = \frac{2095}{450} = \left(\frac{261}{450}\right)^2 = 4.3192$$

$$\sigma_x = d.\sigma_y = 250 \times \sqrt{4.3192} = 250 \times 2.0783 = 519.6$$

Example 10**:** *The score of two batsmen, A and B, in 10 innings are as follows. Find which of the batsmen is more consistent in scoring.* [I.C.W.A., Jan 1970]

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Solution: The coefficient of variation is used to measure dispersion and the batsman with the lowest coefficient of variation is more consistent. Since the values are high, let us change the origin to 50.

Batsman A			Batsman B		
X	$y = x - 50$	y^2	X	$y = x - 50$	y^2
32	-18	324	19	-31	961
28	-22	484	31	-19	361
47	-3	9	48	-2	4
63	13	169	53	3	9
71	21	441	67	17	289
39	-11	121	90	40	1600
10	-40	1600	10	-40	1600
60	10	100	62	12	144
96	46	2116	40	-10	100
14	-36	1296	80	30	900

-	$\sum y$= -40	$\sum y^2 = 6660$	-	$\sum y$= 0	$\sum y^2 = 5968$

For batsman A

Mean = $50 + \frac{-40}{10} = 46$

S.D. = $\sqrt{\frac{6660}{10} - \left(\frac{-40}{10}\right)^2} = \sqrt{650} = 25.5$

C.V. = $\frac{25.5}{46} \times 100 = 55$

For batsman B

Mean = $50 + \frac{0}{10} = 50$

S.D. = $\sqrt{\frac{5968}{10} - \left(\frac{0}{10}\right)^2} = \sqrt{596.8} = 24.4$

C.V. = $\frac{24.4}{50} \times 100 = 49$

Since for batsman B, the coefficient of variation is smaller, he is more consistent.

3.8.4. Merits and demerits of Standard Deviation

Merits

1. Standard deviation is rigidly defined and based on all observations.
2. It is capable of further algebraic treatment and possesses many mathematical properties.
3. The effect of fluctuations in sampling is less in standard deviation than most other measures of dispersion.
4. For comparing variability of two or more series, coefficient of variation is considered as most appropriate and this is based on S.D. and mean.

Demerits

1. It is not easy to understand and to calculate.'
2. It gives more weight to the extremes and less to the values nearer to mean.

3.9. Lorenz Curve

Lorenz curve is a diagrammatic method of studying the dispersion of a group. It was first used to measure the economic inequalities such as distribution of income and wealth between different countries. But today, Lorenz curve is also used in business to study the disparities of the distribution of wages, profits etc. Lorenz curve deals with the cumulative values of the variable and the cumulative frequencies rather than its absolute values and given frequencies. It is, in fact, a cumulative percentage curve combining the percentages of each variable. Let us consider the distribution of income among a certain group of individuals. We derive the curve by measuring percent cumulative frequencies on the horizontal axis and percent cumulative total income on the vertical axis.

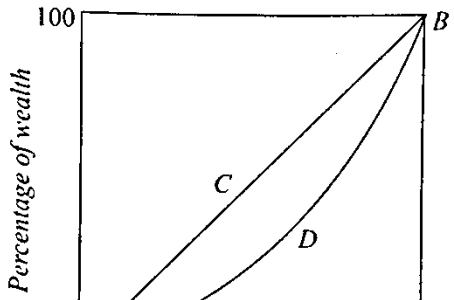


Fig 3.1: Lorenz Curve

We plot the percentages of cumulative values of income against the percentages of cumulative values of frequencies for the given distribution and join these points with a smooth-free hand curve. The curve ADB is a Lorenz curve. The straight line in the figure shows the line of equality and the Lorenz will never cross the line of equal distribution. It will always lie below the line ABC unless the distribution is uniform in which case it will coincide with ABC. The larger the area between the curve ABD and line ABC, the larger is the percentage of poor people and greater is the concentration of wealth in the hands of few. Similarly, if the area is less then there is a low variability in the distribution of income. The Lorenz curve is also known as curve of concentration and the area between the line of equal distribution and the Lorenz curve is known as area of concentration.

3.10. Best Measure of Dispersion: From the above analysis of merits and demerits of each measure of dispersion, we find that the standard deviation has maximum advantages. It almost satisfies all the properties of a good measure of variation. It is based on all values and thus, provides information about the complete series. It can be used in advance statistical calculations like comparison of variability of two data sets. It is very useful in testing hypothesis.

Exercises

1. Find out the range of the following data: [D.S.W., Nov.1973]

Heights (inches)	60-62	63-65	66-68	69-71	72-74
No. of Students	8	27	42	18	5

Ans: 15 inches

2. Calculate the quartile deviation and its coefficient from the following: [C.A., Nov. 1976]

Class Interval	10-15	15-20	20-25	25-30	30-40	40-50	50-60	60-70	Total
Frequency	4	12	16	22	10	8	6	4	82

Ans: 8.05, 30.

3. Calculate the mean deviation from the following data, relating to heights (to the nearest inch) of 100 children:

[I.C.W.A., Jan 1973]

Height (Inches)	60	61	62	63	64	65	66	67	68
No. of Children	2	0	15	29	25	12	10	4	3

Ans: 1.24 inches.

4. The mean and the standard deviation of a sample of size 10 were found to be 9.5 and 2.5 respectively. Later on, an additional observation became available. This was 15.0 and was included in the original sample. Find the mean and the standard deviation of the 11 observations. [I.C.W.A., June 1975]

5. Find the standard deviation for the distribution given below: [Dip. Management 1967]

x	1	2	3	4	5	6	7
Frequency	10	2	30	35	14	10	2

Ans: 1.4

6. Find the coefficient of variation of the following data: [C.U., B.Com. 1986]

Marks	0-14	10-20	20-30	30-40	40-50
No. of Students	4	10	16	12	8

Ans: 43.2%

7. Compute the arithmetic mean, standard deviation and mean deviation above the mean for the following data: [I.C.W.A., Dec. 1978]

Scores	4-5	6-7	8-9	10-11	12-13	14-15	Total
Frequency	4	10	20	15	8	3	60

Ans: 9.23; 2.48; 2.03.

8. Out of 400 observations, 100 observations have the value one and the rest of the observations are zero. Find the mean and s.d. of 400 observations together. [B.U., B.A. (Econ) 1966]

9. From the prices of shares X and Y below find out which is more stable in value.

[I.C.W.A., Dec. 1976- old]

X	35	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

Ans: 11.6; 1.9

10. The runs scored by cricketers A and B during a 10 consecutive innings are shown below, for which of the cricketers is more consistent in scoring runs

[C.U., B.Com. (new) 2006]

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Ans: C.V. of A = 55%; C.V. of B = 49%

UNIT IV **CORRELATION AND REGRESSION**

5.1. Objectives: - The objective of this chapter is to know, how in real world one variable can be affected by another variable. How we can estimate the value of one unknown variable with the help of a known variable.

5.2. Introduction: So far we have considered only one variable. Its mean, dispersion and skewness and so on. This is a univariate analysis. In practical life we deal with two or more than two variables and we may find some degree of association between these variables. Suppose let us consider the income and expenditure of a family, we know the expenditure is directly related to income which means as income of a family rises, their expenditure also increases. Thus, these two variables are associated positively. Likewise, the amount of rainfall and the yield of a certain crop, height and weight of a group of individuals, education qualification and getting a job, inflation in an economy and purchasing power of the consumers etc., has some degree of association. In order to learn how much is the magnitude of this degree of association, the concept of correlation is used.

In some cases, there may be one variable of particular interest and where other variables are regarded as auxiliary variable which helps to find the value of the former variable. In these cases one is interested in estimating an equation which would predict the value of the variable in interest. In this case the concept of regression is very

useful and the equation is known as regression equation. In this chapter we will discuss how two variables are related and how some variables help to find the value of one variable.

5.3. Correlation:

Correlation gives the degree of association between two or more variable. In this chapter we are interested in finding the correlation between two variables only. Since two variables are being considered, here the correlation is the analysis of bivariate data. Suppose there are two variables x and y which are related to in such a way that variation in the magnitude of one variable causes variation in the magnitude of the other variable, then they are said to be correlated. If y tends to increase as x increases, the variables are said to be positively correlated. For example income and expenditure of a family. If y tends to fall as x rises, the variables are negatively correlated (inflation and purchasing power of a consumer) and if the values of y are not affected by change in the values of x, then the variables are said to be uncorrelated. If the magnitude of change in one variable tends to bear a constant ratio to the amount of change in other variable, the correlation is said to be a linear.

In this section we shall discuss only linear correlation or simple correlation. This is measured by correlation coefficient. Before going into detail about correlation let us first define a bivariate data.

5.3.1. Bivariate data: The word bi means two, which implies bivariate means two variables. Suppose, we are collecting data on two characters of a group of individuals. Then these two characters are considered as two variables. The data which are related to the simultaneous measurement of two variables are called bivariate data. Correlation is measurement of degree of association between two such variables.

5.3.2. Measurement of Correlation:

Correlation of a set of data of two variables can be measured in two ways.

I. Scatter Diagram

II. Correlation coefficient

I. Scatter diagram: Scatter diagram is a graphical representation of two variables relating to simultaneous measurement. Each pair of observation can be represented by a point on the graph paper- value of one variable on X-axis and another on Y-axis. If there are n numbers of observations, then we shall have n points. Scatter diagram shows the degree of association between two variables or the correlation between them. In the following figures we will understand how scatter diagram helps to find the degree of association between variables.

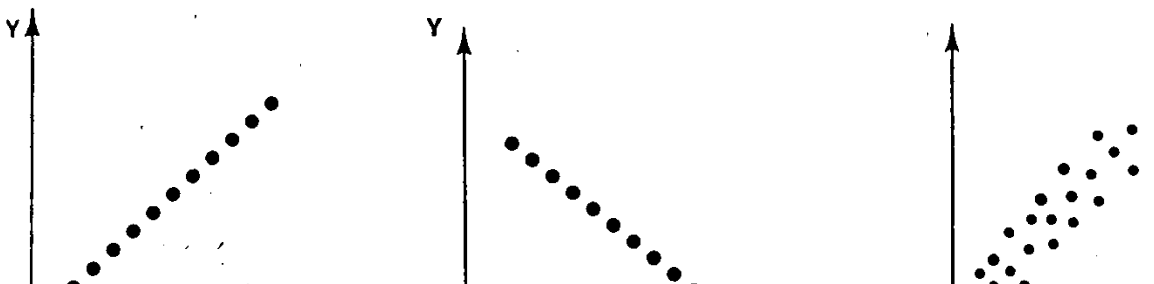
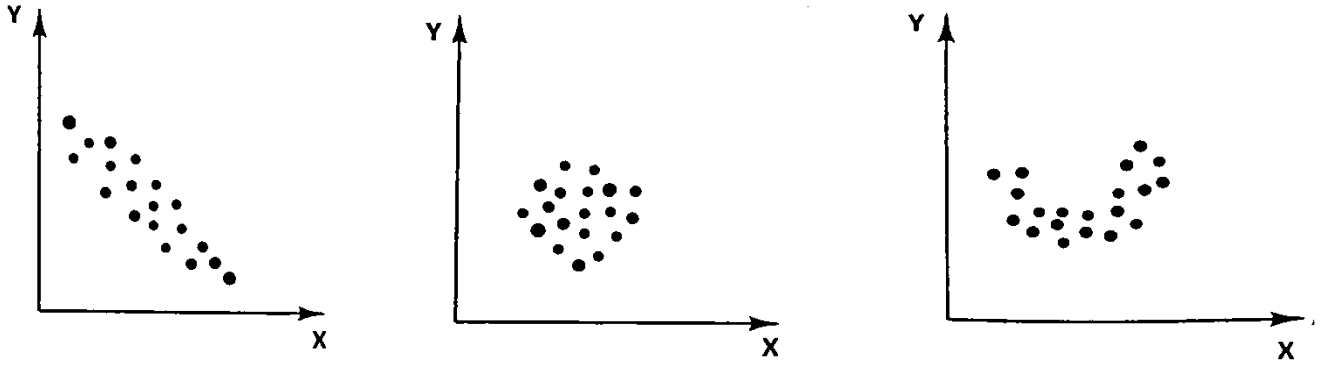


Fig 5.1 (a) +1(Perfect Correlation) (b) -1 (c) Positive



(d) Negative (e) Zero (f) Zero

In the above figures, figure (a) and (b) shows perfect correlation ($+1$ or -1) indicating a straight upward or downward sloping line according to the slope of line positive or negative. Figure (c) shows a positive relation between X and Y variables where the points are scattered on the graph from bottom left hand corner to the top right. In other words, the association between two variables is direct. In figure (d), the dots on graph are spread from upper left corner to bottom right, which means for high value of one variable is associated with low value of other. In this case the correlation is negative. From figure (e) and (f) we are not able to get a straight line but is a concentration of a group of points. In this case no exact association can be found between the variables. Thus the correlation is zero.

II. Correlation Coefficient: Correlation coefficient or coefficient of correlation is used to find the correlation between two variables using statistics.

Definition: Let $(x_1,\ y_1), (x_2,\ y_2), \dots\dots\dots (x_n,\ y_n)$ be a set of observations on two variables x and y. The correlation coefficient (denoted by r) between variables x and y is defined as

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Where σ_x and σ_y are standard deviations of x and y respectively and $cov(x,y)$ is covariance of x and y. This expression is known as Pearson's product moment formula. If $(x_1,\ y_1), (x_2,\ y_2), \dots\dots\dots (x_n,\ y_n)$ be a set of observations on two variables x and y, then covariance is given by the formula

$$cov(x,y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) \qquad (1)$$

Expanding equation 1 on the right side we get,

$$cov(x,y) = \frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right) = \sum xy - n\bar{x}\bar{y}$$

$$var\,(x)\ or\ {\sigma_x}^2 = \frac{1}{n}\sum(x-\bar{x})^2 \quad \therefore\ \sigma_x = \sqrt{\frac{1}{n}\sum(x-\bar{x})^2}$$

$$var\,(y)\ or\ {\sigma_y}^2 = \frac{1}{n}\sum(y-\bar{y})^2 \qquad \therefore \sigma_y = \sqrt{\frac{1}{n}\sum(y-\bar{y})^2}$$

$$\therefore r = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\sum(x-\bar{x})^2\frac{1}{n}\sum(y-\bar{y})^2}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \qquad (2)$$

Expanding the expression (2), we get

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{[(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)]}}$$

Multiplying the numerator and denominator by *n*, and since $n\bar{x} = \sum x\ and\ n\bar{y} = \sum y$, we may write

$$r = \frac{n\sum xy - \sum x\sum y}{\sqrt{[\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}]}}$$

5.3.3. Properties of Correlation Coefficient

1. The correlation coefficient is independent of both origin and scale.

Proof: Let $u = \frac{x-a}{c}\ and\ v = \frac{y-b}{d}$ where a, b, c and d are arbitrary constants and c and d are positive.

$$u = \frac{x-a}{c} \qquad \text{and} \qquad v = \frac{y-b}{d}$$

$$\Rightarrow x = a + cu \qquad \text{and} \qquad \Rightarrow y = b + dv \qquad (i)$$

Summing both sides and dividing by *n*, we get

$$\bar{x} = a + c\bar{u} \qquad \text{and} \qquad \bar{y} = b + d\bar{v} \qquad (ii)$$

Subtracting (ii) from (i), we get

$$(x - \bar{x}) = c(u - \bar{u}) \qquad \text{and} \qquad (y - \bar{y}) = d(v - \bar{v})$$

Substituting these values in 2, we get

$$r_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{cd\sum(u-\bar{u})(v-\bar{v})}{cd\sqrt{\sum(u-\bar{u})^2\sum(v-\bar{v})^2}}$$

$$= \frac{\sum(u-\bar{u})(v-\bar{v})}{\sqrt{\sum(u-\bar{u})^2\sum(v-\bar{v})^2}} = r_{uv}$$

Thus, $r_{xy} = r_{uv}$. So, correlation coefficient is independent of both origin and scale.

2. The correlation coefficient lies between -1 *and +1. That is r cannot exceed one numerically.*

Proof: In the expression 2 let us write,

$$u_i = \frac{x_i - \bar{x}}{\sigma_x} \qquad \text{and} \qquad v_i = \frac{y_i - \bar{y}}{\sigma_y}$$

Then,

$$\sum u_i = \sum\left(\frac{x_i-\bar{x}}{\sigma_x}\right)^2 = \frac{\sum(x_i-\bar{x})^2}{{\sigma_x}^2} = n.\frac{{\sigma_x}^2}{{\sigma_x}^2} = n$$

Similarly,

$$\sum v_i = n$$

Again,

$$\sum u_i v_i = \sum\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} = \frac{n.cov(x,y)}{\sigma_x\sigma_y} = nr$$

Where r is correlation coefficient.

Since, $(u_i + v_i)^2$ is always positive, the sum of all such squares will also be positive.

$$\sum(u_i+v_i)^2 \geq 0$$

$$or, \quad \sum {u_i}^2 + \sum {v_i}^2 + 2.\sum u_i v_i \geq 0$$

$$or, \quad n + n + 2nr \geq 0$$

$$or, \quad 2n + 2nr \geq 0$$

$$or, \quad 2n(1+r) \geq 0$$

$$or, \quad r \geq -1; i.e \quad -1 \leq r$$

Similarly, Since $(u_i - v_i)^2$ cannot be negative,

$$\sum(u_i-v_i)^2 \geq 0$$

$$or, \quad \sum {u_i}^2 + \sum {v_i}^2 - 2.\sum u_i v_i \geq 0$$

$$or, \quad n + n - 2nr \geq 0$$

$$or, \quad 2n - 2nr \geq 0$$

$$or, \quad 2n(1 - r) \geq 0$$

$$or, \quad 1 \geq r; \ i.e \quad r \leq 1$$

Combining the result, we have $-1 \leq r \leq 1$.

Note: Since standard deviation can never be negative, the negative value of *r* is only possible because of negative covariance. Therefore covariance can be negative.

3. The correlation coefficient is a pure number, that is, it is independent of units of measurement x and y.

4. If two variables are independent of each other, then the value of r is zero, since the variables are unrelated, so r is zero.

Example 1: *Find the coefficient of correlation from the following data.* [C.U., B.A (Econ) '69]

X	65	63	67	64	68	62	70	66
y	68	66	68	65	69	66	68	65

Solution: Since the correlation coefficient is unaffected by the change in origin and scale, let us change the origin of *x* and *y* to 65 and 67 respectively.

X	y	$u = x - 65$	$v = y - 67$	u^2	v^2	uv
65	68	0	1	0	1	0
63	66	-2	-1	4	1	2
67	68	2	1	4	1	2
64	65	-1	-2	1	4	2
68	69	3	2	9	4	6
62	66	-3	-1	9	1	3
70	68	5	1	25	1	5
66	65	1	-2	1	4	-2
$\sum x = 525$	$\sum y = 535$	$\sum u = 5$	$\sum v = -1$	$\sum u^2 = 53$	$\sum v^2 = 17$	$\sum uv = 18$

$$\sigma_u{}^2 = \frac{\sum u^2}{n} - \left(\frac{\sum u}{n}\right)^2 = \frac{53}{8} - \left(\frac{5}{8}\right)^2 = \frac{399}{64}$$

$$\sigma_v{}^2 = \frac{\sum v^2}{n} - \left(\frac{\sum v}{n}\right)^2 = \frac{17}{8} - \left(\frac{-1}{8}\right)^2 = \frac{135}{64}$$

$$cov(u,v) = \frac{\sum uv}{n} - \left(\frac{\sum u}{n}\right)\left(\frac{\sum v}{n}\right) = \frac{18}{8} - \left(\frac{5}{8}\right)\left(-\frac{1}{8}\right) = \frac{149}{64}$$

$$r = \frac{cov(u,v)}{\sigma_u \sigma_v} = \frac{\frac{149}{64}}{\sqrt{\frac{399}{64}}\sqrt{\frac{135}{64}}} = \frac{149}{\sqrt{399 \times 135}} = 0.64$$

5.3.4. Rank correlation:

There are some attributes which are not quantifiable such as intelligence, beauty, honesty, hard word etc. In such cases the correlation coefficient cannot be determined. Therefore, these attributes of individuals in the group can be arranged in order and hence obtaining for each individual a number indicating the rank in the group. The correlation coefficient between the two series of ranks is called Rank Correlation Coefficient. It is given by the formula

$$R = 1 - \frac{6\sum d^2}{n^3 - n}$$

Where *d* represents the difference of the ranks of an individual in the two characters and *n* is the number of individuals. This formula was given by Spearman, and hence is also known as Speraman's formula for rank correlation coefficient.

The value of rank correlation coefficient lies between -1 and +1.

$$-1 \leq R \leq +1$$

The value of *R* will be equal to +1 when the ranks in the two characters are equal and in the same direction and the value is -1 when the ranks are just opposite.

If some individuals have same score in any character, then they are allotted with same rank which is known as tied ranks. In these cases each individual will be allotted average rank. Since the ranking of individuals are changed a small modification is made on the original rank correlation formula to accommodate this change. The modified formula for rank correlation coefficient is given as

$$R' = 1 - \frac{6\left\{\sum d^2 + \frac{\sum (m^3 - m)}{12}\right\}}{n^3 - n}$$

Where *m* denotes the number of individuals involved in a tie whether in the first or second series.

Example 2: *In a contest, two judges ranked eight candidates in order of their preferences, as shown in the following table. Find the rank correlation coefficient.* [I.C.W.A., June '75]

Candidates	A	B	C	D	E	F	G	H
1st Judge	5	2	8	1	4	6	3	7
2nd Judge	4	5	7	3	2	8	1	6

Solution:

Candidates	Ranks by		$d = x - y$	d^2
	Judge 1 x	Judge 2 Y		
		4		
A	5	5	1	1
B	2	7	-3	9
C	8	3	1	1
D	1	2	-2	4
E	4	8	2	4
F	6	1	2	4
G	3	6	2	4
H	7		1	1
n = 8	-	-	-	$\sum d^2$ =28

$$R = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \times 28}{8^3 - 8} = 0.67$$

Example 3: *Compute the rank correlation coefficient between X and Y from the data given below.*

X	8	10	7	15	3	20	21	5	10	14	8	16	22	19	6
Y	3	12	8	13	20	9	14	11	4	16	15	10	18	23	25

Solution: Since the observation 8 is repeated twice, the rank given is the average of 5^{th} and 6^{th} and observation 10 is also repeated twice, its rank is the average of 7^{th} and 8^{th}.

X	Rank of X	Y	Rank of Y	d = Rank(x)-Rank(y)	d^2
8	5.5	3	1	4.5	20.25
10	7.5	12	7	0.5	0.25
7	4	8	3	1	1
15	10	13	8	2	4
3	1	20	13	-12	144
20	13	9	4	9	81
21	14	14	9	5	25
5	2	11	6	-4	16
10	7.5	4	2	5.5	30.25
14	9	16	11	-2	4
8	5.5	15	10	4.5	20.25
16	11	10	5	6	36
22	15	18	12	3	9
19	12	23	14	-2	4
6	3	25	15	-12	144
-	-	-	-	-	$\sum d^2 = 539$

$$R' = 1 - \frac{6\left\{\sum d^2 + \frac{\sum(m^3 - m)}{12}\right\}}{n^3 - n}$$

Since item 8 and 10 are repeated twice in X-series then $m = 2$ for each item.

$$= 1 - \frac{6\{539 + \frac{1}{12}(2^2 - 2) + \frac{1}{12}(2^2 - 2)}{15^3 - 15}$$

$$= 1 - \frac{6(539 + 0.5 + 0.5)}{3360} = 1 - \frac{3240}{3360} = 0.0357$$

5.4. Regression:

In a bivariate data analysis, when the value of unknown variable can be determined with the help of the value of known variable, then we use the concept of regression. In regression we estimate or predict the value of unknown dependent variable by using the value of known independent variable. In this case, we find two equations known as regression equations. Suppose we have variables *x* and *y*, then the equation $y = a + bx$ will be regression equation of *y* on *x* if *y* is unknown and *x* is known. Similarly, $x = a' + b'y$ is the regression equation of *x* on *y* if *x* is unknown and *y* is known. In this section we shall discuss only linear regression. When the relation between two correlated variables is expressed in a straight line, then it is known as linear regression.

The simplest relationship between *y* and *x* could perhaps be a linear deterministic function given by

$$y_i = a + bx_i \qquad (1)$$

In the above equation x is the independent variable or explanatory variable and y is the dependent variable or explained variable. The subscript *i* represents the observation number, *i* ranges from 1 to *n*.

For simplification we write equation (1) as $y = a + bx$

5.4.1. Regression Lines: The regression lines give the best estimate of one variable for any given value of the other. If there are two variables, we shall have two regression lines; one of *y* on *x* and other *x* on *y*. Equation (1) represents the regression line of *y* on *x* which gives the best estimate for the values of *y* for any specified value of *x*.

Similarly, the equation $x = a + by$ represents the regression line of *x* on *y* which gives the best estimate for the values of *x* for any specified value of *y*.

We assume the relationship between *y* and *x* to be stochastic and add one error term in (1). Thus our stochastic model is $y = a + bx + e \qquad (2)$

where *e*, is the error term. In real life situations *e* represents randomness in human behaviour and excluded variables, if any, in the model. Remember that the right hand side of (2) has two parts, viz., i) deterministic part (that is, $a + bx$), and ii) stochastic or randomness part (that is, *e*,). Equation (2) implies that even if *x*, remains the same for two observations, *y*, need not be the same because of different *e*. Thus, if we plot (2) on a graph paper the observations will not remain on a straight line.

Suppose we have data on rainfall and agricultural production for ten years as shown in the table

Rainfall (mm)	61	61	64	70	72	76	80	84	87	91
Agricultural Production (tons)	34	36	37	42	42	46	49	52	55	57

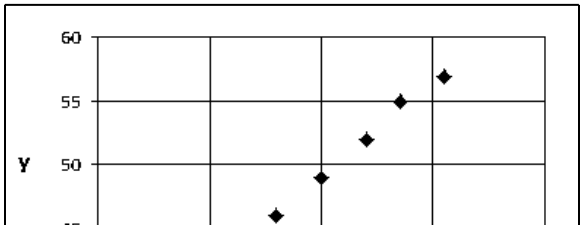


Fig 5.2: Scatter Plot

Fig 5.2: Scatter Plot

We plot the data on a graph paper. The scatter diagram looks something like Fig. 5.2. We observe from Fig. 5.2 that the points do not lie strictly on a straight line. But they show an upward rising tendency where a straight line can be fitted. Let us draw the regression line along with the scatter plot.

In the figure 5.3 the vertical difference between the regression line and the observations is the error e. The value corresponding to the regression line is called the predicted value or the expected value. On the other hand, the actual value of the dependent variable corresponding to a particular value of the independent variable is called the observed value. Thus 'error' is the difference between predicted value and observed value.

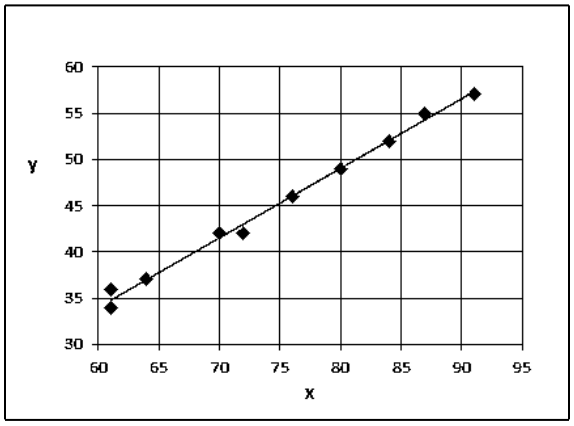


Fig 5.3: Regression Line

5.4.2. Regression Equations: The regression equation can be represented in different form and is derived using the method of least squares. Let us derive these two lines one by one.

(i) Regression equation of *y* on *x*

The regression equation *y* on *x* is the equation of the best fitting straight line in the form of $y = a + bx$, obtained by the method of least squares.

Let $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n,\ y_n)$ be a set of *n* pairs of observations, let us fit a straight line of the form

$$y = a + bx \qquad (i)$$

Applying the method of least squares, the constants *a* and *b* are obtained by solving the normal equations

Summing the equation (i), we get

$$\sum y = na + b\sum x \qquad (ii)$$

Multiplying both sides by *x*, we get

$$\sum xy = a\sum x + b\sum x^2 \qquad (ii)$$

Dividing both side of (i) by *n*, we get

$$\bar{y} = a + b\bar{x} \qquad so\ that \qquad a = \bar{y} - b\bar{x}$$

Substituting this in equation (i), we get

$$y - \bar{y} = b(x - \bar{x}) \qquad (iv)$$

Again multiplying (ii) by $\sum x$ and (iii) by *n*, we get

$$\sum x \sum y = a + b(\sum x)^2$$

$$n\sum xy = na\sum x + nb\sum x^2$$

Subtracting the first from second, we get

$$n\sum xy - \sum x\sum y = b\{n\sum x^2 - (\sum x)^2\}$$

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2}$$

Dividing both numerator and denominator by n^2, we get

$$b = \frac{\sum xy/n - \left(\sum x/n\right)\left(\sum y/n\right)}{\sum x^2/n - \left(\sum x/n\right)^2} = \frac{cov(x,y)}{{\sigma_x}^2}$$

Writing *b* with usual subscripts, we have from (iv)

$y - \bar{y} = b_{yx}(x - \bar{x})$, where $b_{yx} = \frac{cov(x,y)}{{\sigma_x}^2}$. This is the required equation of *y* on *x*. Since

$r = cov(x,y)/\sigma_x\sigma_y$, we see that $r\sigma_x\sigma_y = cov(x,y)$

Substituting this we get,

$$b_{yx} = \frac{cov(x,y)}{{\sigma_x}^2} = r\,\frac{\sigma_y}{\sigma_x}$$

where b_{yx} is known as Regression Coefficient of *y* on *x*.

(ii) Regression equation of *x* on *y*

Similarly, following the steps above, we get the regression equation of *x* on *y* as

$$x - \bar{x} = b_{xy}(y - \bar{y}), \text{ where } b_{xy} = \frac{cov(x,y)}{\sigma_y{}^2} = r\,\frac{\sigma_x}{\sigma_y}$$

and b_{xy} is the Regression Coefficient of *x* on *y*.

5.4.3. Properties of linear Regression

1. The product of two regression coefficient is equal to the square of correlation coefficient. That is

$$r^2 = b_{yx}\,.b_{xy} \qquad \Rightarrow \quad r = \pm\sqrt{b_{yx}\,.b_{xy}}$$

The sign of correlation coefficient is same as that of regression coefficients. If regression coefficients are positive, \r is positive and if regression coefficients are negative r is negative.

2. If one of the regression coefficients is greater than unity, other must be less than unity.

3. Regression coefficients are independent of change of origin but not of scale.

Note:

i. The two lines will coincide when the value of correlation coefficient is either +1 or -1.
ii. The two lines will be perpendicular to each other when value of correlation coefficient is zero.
iii. The smaller is the angle between the two regression lines, greater is the degree of correlation.
iv. The two regression lines will intersect at $\bar{x}$ and $\bar{y}$.

Example 4: From the following data, obtain the two regression equations. [C.A., May '77]

Sales	91	97	108	121	67	124	51	73	111	57
Purchases	71	75	69	97	70	91	39	61	80	47

Solution: Let us denote sales by the variable *x* and purchases by *y*.

x	Y	$u = x - \bar{x}$	$v = y - \bar{y}$	u^2	v^2	*uv*
91	71	1	1	1	1	1
97	75	7	5	49	25	35

108	69	18	-1	324	1	-18
121	97	31	27	961	729	837
67	70	-23	0	529	0	0
124	91	34	21	1156	441	714
51	39	-39	-31	1521	961	1209
73	61	-17	-9	289	81	153
111	80	21	10	441	100	210
57	47	-33	-23	1089	529	759
$\sum x = 900$	$\sum y = 700$	$\sum u = 0$	$\sum v = 0$	$\sum u^2 = 6360$	$\sum v^2 = 2868$	$\sum uv = 3900$

From the above table we have,

$$\bar{x} = \frac{\sum x}{n} = \frac{900}{10} = 90 \qquad and \qquad \bar{y} = \frac{\sum y}{n} = \frac{700}{10} = 70$$

$$b_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum uv}{\sum u^2} = \frac{3900}{6360} = 0.6132$$

$$b_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{\sum uv}{\sum v^2} = \frac{3900}{2868} = 1.361$$

Equation of line of regression of *y* on *x*

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

$$\Rightarrow y - 70 = 0.6132(x - 90)$$

$$\Rightarrow y = 0.6132x + 14.812$$

Equation of line of regression of *x* on *y*

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

$$\Rightarrow x - 90 = 1.361(y - 70)$$

$$\Rightarrow x = 1.361y - 5.27$$

Example 5: Find the regression equations from the following data and hence find the value of correlation coefficient. Also find the value of *y* when *x* = 80. [C.U., B.Com. (new) 2005]

X	41	82	62	37	58	96	127	74	123	100
Y	28	56	35	17	42	85	105	61	98	73

Solution:

x	Y	$u = x - \bar{x}$	$v = y - \bar{y}$	u^2	v^2	uv
41	28	-39	-32	1521	1024	1248
82	56	2	-4	4	16	-8
62	35	-18	-25	324	625	450
37	17	-43	-43	1849	1849	1849
58	42	-22	-18	484	324	396
96	85	16	25	256	625	400
127	105	47	45	2209	2025	2115
74	61	-6	1	36	1	-6
123	98	43	38	1849	1444	1634
100	73	20	13	400	169	260
$\sum x = 800$	$\sum y = 600$	$\sum u = 0$	$\sum v = 0$	$\sum u^2 = 8932$	$\sum v^2 = 8102$	$\sum uv = 8338$

From the above table, we have

$$\bar{x} = \frac{\sum x}{n} = \frac{800}{10} = 80 \qquad and \qquad \bar{y} = \frac{\sum y}{n} = \frac{600}{10} = 60$$

$$b_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum uv}{\sum u^2} = \frac{8338}{8932} = 0.933$$

$$b_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{\sum uv}{\sum v^2} = \frac{8338}{8102} = 1.03 \ (approx)$$

Equation of line of regression of y on x

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

$$\Rightarrow y - 60 = 0.933(x - 80)$$

$$\Rightarrow y = 0.933x - 14.64$$

Equation of line of regression of x on y

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

$$\Rightarrow x - 80 = 1.03(y - 60)$$

$$\Rightarrow x = 1.03y + 18.2$$

$$r = \sqrt{b_{yx} \, . \, b_{xy}} = \sqrt{0.933 \times 1.03} = +0.98$$

For $x = 80, \quad y = 0.933 \times 80 - 14.64 = 60.$

5.4.4. Difference between Correlation and Regression Analysis

1. Correlation analysis is used to measure the co-relationship or the association between the two variables. Whereas regression analysis is used to measure value of one variable on the basis of given values of another variable.
2. Correlation analysis represents the linear relationship between two variables while regression analysis explains how to fit a best line and estimate one variable on the basis of another variable.
3. There is no difference between dependent and independent variable in correlation analysis whereas there is a major difference between dependent and independent variable in the analysis of regression.
4. Correlation indicates the degree of association between variables. As opposed to, regression reflects the impact of the unit change in the independent variable due to unit change in the dependent variable.
5. Correlation aims at finding a numerical value that expresses the relationship between variables. Unlike regression whose goal is to predict values of the random variable on the basis of the values of fixed variable.

Exercises

1. Marks of 10 students of Mathematics and Statistics are given below; calculate product-moment correlation coefficient. [I.C.W.A., June 1974]

Mathematics	32	38	48	43	40	22	41	69	35	64
Statistics	30	31	38	43	33	11	27	76	40	59

Ans: $r = 0.94$.

2. Ranking of 10 trainees at the beginning (x) and at the end (y) of a certain course are given below; calculate spearman's rank correlation coefficient. [I.C.W.A., June 1995]

Trainees	A	B	C	D	E	F	G	H	I	J
X	1	6	3	9	5	2	7	10	8	4
Y	6	8	3	7	2	1	5	9	4	10

Ans: $R = 0.394$

3. The competitors in a musical contest were ranked by 3 three judges A, B and C in the following order:

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9

Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music. [I.C.W.A., Dec. 1978]

Ans: R_{AB} = -0.21; R_{BC} = -0.30; R_{AC} = +0.64.

4. Find the regression equation of *y* on *x* where *x* and *y* are the marks obtained by 10 students as given below: [C.A., (Foundation). May 2002]

x	20	60	55	45	75	35	25	90	10	50
y	20	45	65	40	55	35	15	80	25	50

Ans: b_{xy} = 1.105; $y = 1.105x - 1.015$

UNIT V TESTING OF HYPOTHESIS

5.1. Objectives: The objectives of this chapter are to know what we mean by testing of hypothesis and how this hypothesis is being setup. To understand what are null and alternative hypotheses. Further we will understand what the different types of error are and different types of test.

5.2. Introduction: In unit 1, we have discussed how the statistics of samples being used to draw conclusion regarding the characteristics of population parameter. But to find if these sample statistics are really relevant or not, we use a method of testing these statistics.

Suppose, we are required to test the effectiveness of a new drug in curing a flu, we need not to take each and every person to whom it have been injected but a representative sample would be enough and test whether the new drug is more effective than existing drug. Similarly, it may be required to decide whether the population from which the sample has been obtained is normal distribution with mean = μ_0 and standard deviation = σ. In order to reach such conclusion we first need to make certain assumption about the characteristics of the population, particularly, about its probability distribution or the values of its parameters. These kinds of assumptions are known as ***statistical hypothesis***. Hypothesis is a statement or assertion or claim about the population parameter. The process which help us to decide whether a certain hypothesis is true or not, is known as ***Test of Hypothesis*** or ***Test of Significance***.

Thus, the testing of hypothesis is a very important aspect of sampling theory. It enables us to decide on the basis of sample observations, the deviation of an observed sample statistic from the population parameter and the

deviation between two independent sample statistics is significant or might be attributed to the fluctuations of sampling.

5.3. Some Concepts

1. **Null Hypothesis:** Null hypothesis is statement about the value of a population parameter and it is denoted by H_0. It is usually a hypothesis of no difference. From the above example we can write the null hypothesis as

$$H_0: \mu = \mu_0$$

The above H_0 means the population mean is as same as μ_0. That there is no difference between these two. According to Prof. A.R. Fisher: "*Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true."*

2. **Alternative Hypothesis:** Alternative hypothesis is a statement that is accepted if evidence proves null hypothesis to be false and it is denoted by H_1. The null hypothesis is tested against an alternative hypothesis which in the above case, may be either that the population mean is not μ_0, or that it is greater than μ_0, or that it is less than μ_0; i.e., any one of the following is possible.

$$H_1: \mu \neq \mu_0, \qquad H_1: \mu > \mu_0, \qquad H_1: \mu < \mu_0$$

We have to keep in mind that null hypothesis and alternative hypothesis are mutually exclusive, that is, both cannot be true simultaneously. Secondly, both H_0 and H_1 exhaust all possible options regarding the parameter, that is, there cannot be a third possibility.

The sample is then analyzed to decide whether to reject or not to reject the null hypothesis H_0. It is a rare coincidence that sample mean $(\bar{x})$ is equal to population mean (μ). In most cases we find a difference between $(\bar{x})$ and (μ). Is the difference because of sampling fluctuation or is there a genuine difference between the sample and the population? In order to answer this question we need a test statistic to test the difference between the two. If the difference is large after testing, the null hypothesis is rejected, and we question the validity of our assumption. If the difference is not so large, H_0 is not rejected, and the difference may be considered to have arisen solely due to fluctuations of sampling.

3. **Test Statistic:** The test static is a mathematical formula that us to decide the likelihood of obtaining sample outcomes if the null hypothesis is true. The value of test statistics determines the final decision regarding acceptance or rejection of null hypothesis. For large sample (n>30), the standard normal variable corresponding to the statistic $\bar{x}$ is given by

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

asymptotically as $n \to \infty$. The value of Z under the null hypothesis is known as test statistic.
If the value of teat statistic falls in the critical region, the null hypothesis is rejected.

4. **Critical Region:** Suppose for each sample of a population, a statistic $\bar{x}$ is calculated. Then we will get a series of values of these statistics. Each of these values may be used to test some null hypothesis. Some values may lead to rejection of null hypothesis while some may lead to acceptance of null hypothesis. These sample statistics may be divided into two disjoint groups, one leading to rejection of H_0 and the other leading to acceptance of H_0. The statistics which lead to rejection of H_0 gives us a region called Critical Region or Rejection Region while those which lead to the acceptance of H_0 gives us a region called Acceptance Region.
5. **Type I and Type II Errors:** While testing the hypothesis there are some error associated with it. There are two types of errors we find after examining the sample from population. These are Type I Error and Type II Error.

   Type I Error: The error of rejecting H_0, when H_0 is true is known as Type I Error.

   Type II Error: The error of accepting H_0 when actually H_0 is false i.e., accept H_0 when H_1 is true is known as Type II Error.

   If we write, P [Reject Ho when it is true] = P [Type I Error] = α

   And P [Accept Ho when it is wrong] = P [Type II Error] = β

   Then α *and* β are called the sizes of type I error and type I1 error respectively. In practice, type I error amounts to rejecting a lot when it is good and type I1 error may be regarded as accepting the lot when it is bad.

   Thus, P [Reject a lot when it is good] = α and P [Accept a lot when it is bad] = β, where α *and* β are referred to as 'Producer's risk' and Consumer's risk' respectively.

   The probability of type I error is necessary for constructing a test of significance. It is in fact the 'size of the critical region'. The probability of type 11 error is used to measure the 'power' of the test in detecting the falsity of the null hypothesis.

6. **Level of Significance:** The maximum size of Type I error, which we are prepared to risk is known as the level of significance. It is usually denoted by α and given by:

   P [Reject Ho when it is true] = α

   Commonly used levels of significance in practice are 5% (0.05) and 1% (0.01). If we take 5% level of significance, it implies that in 5 samples out of 100, we are likely to reject H_0. In other words it implies that we are 95% confident that our decision to reject H_0 is correct. Level of significance is always fixed in advance before collecting the sample information.

7. **Two tailed and one-tailed Tests of Hypothesis:** The null hypothesis for testing the mean of a population is given as

$$H_0: \mu = \mu_0$$

Against the alternative hypothesis

$H_1: \mu \neq \mu_0$ ……. (1) $H_1: \mu > \mu_0$ …… (2) $H_1: \mu < \mu_0$ …… (3)

The alternative hypothesis (2) and (3) are called one tailed test where (2) is Right tailed where the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$ and (3) is left tailed where the critical region lies entirely in the left tail of the sampling distribution $\bar{x}$.

To test the hypothesis $H_0: \mu = \mu_0$ against the alternative hypothesis (1) $H_1: \mu \neq \mu_0$ (which implies $\mu > \mu_0$ *or* $\mu < \mu_0$) is known as two-tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed we apply the two-tailed test and if alternative hypothesis is one-tailed, we apply one-tailed test.

For example, suppose there are two types of drugs available in the market on flu, let μ_1, and μ_1 as their mean of effectiveness. If we want to test the effectiveness of these two drugs differs significantly, then our bull hypothesis is $H_0: \mu_1 = \mu_2$ and the alternative hypothesis will be $H_1: \mu_1 \neq \mu_2$, thus, giving us a two tailed test.

However if we want to test the effectiveness of drug I is higher than the drug II, then we have

$$H_0: \mu_1 = \mu_2 \quad \text{and} \quad H_1: \mu_1 > \mu_2$$

Thus giving us the right-tailed test. Similarly, if we want to test the effectiveness of drug I is less than the drug II, then we have

$$H_0: \mu_1 = \mu_2 \quad \text{and} \quad H_1: \mu_1 < \mu_2$$

Thus giving us the left tailed test.

8. **Critical Value:** The value of test statistic which separates the critical region and the acceptance region is called the critical value. It depends upon
    i. The level of significance used and
    ii. The alternative hypothesis, whether it is two tailed or one tailed.

5.4. Procedure of Testing a Hypothesis

i. Set up the null hypothesis.
ii. Set up the alternative hypothesis.
iii. Level of significance is fixed in advanced, before drawing of the sample.
iv. Apply sample values to z-statistic.

v. Find out from z-table the critical value according to level of significance.
vi. If the value is lower than the critical value do not reject the null hypothesis.
vii. If the value is greater than the critical value reject the null hypothesis and accept the alternative hypothesis.

5.5. Chi-Square (χ^2) Distribution: Let $Z_1, Z_2, \ldots, Z_n$ be n standard normal variables i.e., $Z_i \sim N(0,1), i = 1, 2, 3, \ldots\ldots, n$. Then, sum of squares of these variables, i.e., $\sum_i^n {Z_i}^2$ is said to have a χ^2 distribution with n degrees of freedom. The degree of freedom means the number of free or independent normal variable contained in χ^2 .

The probability density function of χ^2 distribution is given by

$$f(\chi^2) = \frac{1}{2^{n/2}\Gamma\left(n/2\right)} e^{-\chi^2/2} (\chi^2)^{n/2-1}, \qquad where\ 0 < \chi^2 < \infty$$

Characteristics

5. It is continuous extending from 0 to ∞.
6. It is always positively skewed.
7. Its expectation is equal to its degrees of freedom and its variance is twice of its degrees of freedom.
8. It has additive property i.e., if Z_1, and Z_2 are two independent χ^2 variates with degrees of freedom n_1 and n_2 respectively, then $Z_1 + Z_2$ is also a χ^2 variable with degrees of freedom $n_1 + n_2$.

The Chi-square distribution is used is used in both small and large sample test.

5.6. Uses of Chi-square test: It is mainly used in

i. Test for goodness of fit.
ii. Test for independence of attributes.
iii. Test for a specified standard deviation. (Small Sample test)

i. Test for goodness of fit: Chi-square test of goodness of fit' is a very powerful test of testing the significance of the discrepancy between theory and experiment. It is used to decide whether the observations are in good agreement with a hypothetical distribution, i.e., whether the sample may be supposed to have arisen from a specified population.

To test the null hypothesis that there is no significant difference between the observed and the theoretical or hypothetical values, i.e., there is good compatibility between theory and experiment, we use the test statistic

$$\chi^2 = \sum_i^n \left[\frac{(O_i - E_i)^2}{E_i}\right] \qquad (i)$$

Which χ^2 distribution with $(n - 1)$ degrees of freedom where $O_i, (i = 1, 2, \ldots., n)$ is a set of observed (experimental) frequencies and $E_i, i = 1, 2, \ldots., n$ is corresponding set of expected (theoretical or hypothetical) frequencies.

Equation (i) is known as a goodness of fit chi-square. If the observed value of the statistic exceeds the tabulated value of χ^2 at a given level of significance, the null hypothesis is rejected.

Chi-square test is also used to test the independence of attributes and for specified standard deviation.

Self-Assessment:

i. What do you mean by testing of a hypothesis?
ii. What is null hypothesis?
iii. What is an alternative hypothesis?
iv. Define type I and type II error.
v. What are one tailed and two tailed tests?
vi. What do you mean by goodness of fit?