INTRODUCTION

The word statistics seems to have been derived from the Latin word Status or Italian word statista or the German word statistik or the French word statistique.

In ancient times the scope of statistics was primarily limited in keeping the records of the population in regard to age, sex-wise, birth, death, property, wealth etc. of a country. This knowledge was used as a tool to know the manpower and also to fix taxes, and levies. The state or the Government collected statistics for administrative purposes.

The methodologies for statistical analysis have been developed to study natural, social, political or economic phenomenon in a systematic and scientific manner. These methodologies depend on experience study of thing and therefore, are based on experience or empirical investigations. An empirical investigation consists of collecting information through observations and analyzing the information for drawing conclusions.

Some times, we pass judgment on observations and say that they are in the normal range or unusually high or low, a given day's temperature is above normal. We use phrases like average student in a class, the range of marks in a given subject. In all this, we make use of empirical investigations, even though they may be in a very crude form.

Thus, in an empirical investigation facts are recorded through observations. When the information or observations are recorded in quantity, we have quantified the information. For ex, weights, heights, age, no. of accidents , production, sales, etc.The systematic and scientific treatment of quantitative measurement is precisely known as Statistics.

Statistics has been defined in two ways. Some writers define it as 'Statistical data' i.e Numerical statement of facts, while others define it as ' Statistical methods'i.e The complete body of the principles and techniques used in collecting and analyzing such data.

Write down the various definitions of statistics:

Statistics as statistical data:

Bowley defines statistics, as numerical statement of facts in any department of enquiry placed in relation to each other.

Webstar defines statistics as classified facts representing the conditions of the people in a state, especially those facts which can be stated in numbers, or in any other tabular or classified arrangement.

Statistics as statistical methods:

Statistics may be called the science of counting – A.L. Bowley.

Statistics is the science which deals with collection , classification ,and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon.- Lovitt

According to **Croxton and Cowden** 'statistics may be defined as a science of collection, presentation, analysis, and interpretation of numerical data.

According to this definition, there are four stages.

1.Collection of data.

The first step of an investigator is the collection of data. Careful collection is needed, because further analysis is based on this. There are different methods of collection of data (census, sampling, primary, secondary etc) and they must be reliable. If the collected data are faulty, results will also be faulty. There fore , the investigator must take special care in collection.

2.Presentation of data:

The collected data are presented in some systematic order to facilitate statistical analysis. The collected data are presented with the help of tables, graphs, and diagrams.

3. Analysis of data:

The next stage is the analysis of presented data. Analysis includes condensation, summarization, conclusion etc, through the means of measures of central tendencies, dispersion skewness, kurtosis, correlation, regression etc,.

4. Interpretation of data:

The last stage in statistical investigation consists of interpretation , i.e drawing conclusions from the data collected and analysed. The interpretation of data is a difficult task and needs a high degree of skill and experience. Correct interpretation will lead to a valid conclusion of the study.

What do you mean by variable?

In every day life, we come across objects, living beings and phenomena, which vary in a number of ways though they may be belonging to the same general category or class. The characteristic on which individuals or objects differ among themselves is called a variable. Thus speed, height, weight, age, sex are variables.

Types of variables:

All variables can be broadly classified in the following categories

1. Quantitative variables
2. Qualitative variables.

When ever the measurement of a variable is possible on a scale in some appropriate units, it is called a quantitative variable. For ex, age, height, income, speed, weight etc.

On the other hand, the measurements on a quantitative variable are called variates.

What are the Types of Quantitative variable?

Quantitative variables may be further classified as Discrete or Discontinuous variable and Continuous variable.

Discrete variable is one where the values of the variable differ from one another by definite amounts. For ex, the number of children in families, the number of accidents per week can be 0, 1, 2,3,4,5 etc.,

A continuous variable can theoretically assume all values within an interval and as such are divisible into smaller and smaller fractional units. Age, distance, height, weight, etc are some examples of a continuous variable. Obviously, the measurements on a continuous variable can never be exact.

A Qualitative Variable shows variation in objects not in terms of magnitude, but in quality or kind. These qualities are called attributes. A qualitative variable is immeasurable with a scale. Sex, nationality, religion, occupation, marital status, literacy are few examples of a qualitative variable.

Explain about functions of statistics:

The following are some important functions of statistics

1. It simplifies complex data

Statistical methods like averages, totals, percentages etc help in condensing mass of data into a few significant figures so as to make them easily understandable.

2. It presents the facts in a definite form:

Statistics deals with the quantitative statements of facts and thus presents them in a precise and definite form. For ex, statements like ‘ the birth rate in India is decreasing’, ‘the prices of commodities are rising,’ etc, do not convey information in a definite form as they do not involve quantitative statements of facts.

3. It provides a technique of comparison:

Comparison of quantitative facts is also an important function of statistics. For ex, area-wise and period wise comparison of data related to production, sales, export, import, population etc. are helpful for drawing valid conclusions about economic activities.

4. It studies relationship:

Correlation analysis is used to find the functional relationship between two or more variables. The relationship is very helpful in making estimates and forecasting future trends. For ex, the relationship between supply and demand, advertisement and sales etc. provides us the basic tool for planning these activities

5. It helps in forecasting:

Statistics is now an indispensable tool in the analysis of activities relating to business, commerce, and industry. Such an analysis is useful for determining trends in activities related to these areas, which in turn form the basis of estimation and forecasting about the phenomenon under study.

Explain the scope of statistics
SCOPE OF STATISTICS:
How statistics is used in States?

Statistics in States:

Statistics is essential for a country. It supplies essential information to run a government. The aim of every state is to promote the welfare of the people. Different policies of government are based on statistics. Periodical collection of data relating to population, national wealth, agriculture, exports, imports, education, crime etc, Are the main guide lines to the government for a good administration. Therefore statistics are the eyes and ears of the state.

How statistics is used in the field of economics?

Statistics in Economics:

Statistics is an indispensable tool in all the aspects of economic study. The problems in economics cannot be studied without the use of statistics. Statistical data and technique of statistical analysis have proved immensely useful in solving a variety of economic problems, such as wages, prices , demand analysis.

How statistics is used in business?

Statistics in Business:

In business, the manufacturer is always interested to estimate the immediate and future demand of his product. The object can be achieved by properly conducted market survey and research which depend on statistical methods. Statistical concepts and methods are also used in controlling the quality of products to the satisfaction of consumer and the producer.

How statistics is used in Medical ,biological sciences?

Statistics and Medical, Biological and Agricultural Sciences:

In medical sciences one is greatly concerned with the causes and incidence of diseases and the results obtained from the use of medicines and drugs. The effectiveness of a drug or medicine can be tested by using the observations.

Similarly in Biological sciences, the precise role of various factors in the growth and development of plant under study may be important. In agricultural sciences, the precise role of factors, viz manure, seed quality, watering process, rainfall, need to be analyzed for optimizing results. Statistical techniques like Analysis of variance and Design of Experiments are useful for isolating the role of such factors.

Explain about the Limitations of Statistics:

1. Statistics deals with aggregate of items and not with individual:

Statistics is the study of mass data and deals with aggregates or group. In fact data on an item, Considered individually does not constitute statistical data. For ex, the income of a family is Rs 4000 does not Convey statistical meaning while the statement ‘the average income of 50 families is Rs 2050’ conveys statistical sense.

2. Statistics deals only with quantitative data:

Statistics are numerical statement of facts. Such characteristics as can not be expressed in numbers are incapable of statistical analysis. Thus, qualitative characteristics like honesty, efficiency, intelligence, blindness and deafness can not be studied directly. For ex, we may study the intelligence of boys on the basis of the marks obtained by them in an examination.

3. Statistical results are true only on an average:

The conclusions obtained statistically are not universally true; they are true only under certain conditions. This is because statistics as a science is less exact as compared to natural sciences.

4. All the values should not be the same:

The values in statistics have to be different. When the amounts of sales in different periods are considered, they will not be equal. The daily productions in a factory will not be the same. In statistics, the observations differ from one another.

5. Statistics can be misused:

The greatest limitation of statistics is that it is liable to be misused. The misuse of statistics may arise because of several reasons. For ex, if statistical conclusions are based on incomplete information, one may arrive at fallacious conclusions. Statistics are like clay and they can not be molded in any manner so as to establish right or wrong conclusion. It requires experience and skill to draw sensible conclusions from the data; otherwise, there is every likelihood of wrong interpretations. Also statistics can not be used to full advantage in the absence of proper understanding of the subject to which it is applied.

COLLECTION OF DATA:

The basic problem of statistical enquiry is to collect facts and figures relating to a particular phenomenon under study, whether the enquiry is in business, economics or social science. The investigator is the person who conducts the statistical enquiry. He is a trained and efficient statistician. He counts or measures the characteristics under study for further statistical analysis. The respondents (informants) are the persons from whom the information is collected. The statistical units are the items on which the measurement is taken.

Collection of data is the process of enumeration together with the proper recording of results. The success of an enquiry is based upon the proper collection of data.

What is meant by primary and secondary data?
Primary and Secondary Data:

What is meant by primary data?
Primary data are those, which are collected for the first time, and they are original in character.

What is meant by secondary data?
Secondary data are those, which are already collected by some one for some purpose and are available for the present study.

For ex, the data collected during census operations are primary data to the department of census and the same data, if used by a research worker for some study is secondary data.

Explain various methods adopted for collecting primary data.
METHODS OF COLLECTING PRIMARY DATA:

1. Direct personal interview:

Under this method of collecting data, the investigator should contact the persons from whom the informations obtained. The investigator must be tactful and courteous in behavior. He asks the questions the informant and collects necessary information. For ex, if one wants to study the living conditions of the people in a village, he has to go to the village, contact the people and get the needed information.

Merits:

1. Original (first hand information) data are collected.
2. True and reliable data can be had
3. The investigator can extract correct information.
4. A high degree of accuracy can be aimed.
5. Uniformity and homogeneity can be maintained.

Demerits:

1. It is unsuitable where the area is large.
2. It is expensive
3. The chances of bias are more
4. An untrained investigator will not bring good result.

2. Indirect oral investigation:

This is a method of collecting primary data through indirect sources. In such
Cases, the investigator interviews the people, who are directly or indirectly connected with the problem under study. This method is usually adopted by enquiry committees or commissions appointed by the government, private bodies etc.

Merits:

1. It is simple and convenient.
2. It saves time, money and labour
3. The information is unbiased
4. Adequate information can be had
5. It can be used in the investigation of a large area.

Demerits:

1. Absence of direct contact is there, the information can not be relied.
2. Interview with an improper man will spoil the result.
3. In order to get the real position, a sufficient number of persons are to be interviewed.
4. The careless attitude of the informant will affect the degree of accuracy.

3. Information from correspondents:

The correspondents gather information on the subject of enquiry and pass on the same to the investigator. This method is adopted by newspaper and journals etc, when information is needed in different fields for ex, accidents, share markets, politics, strikes etc,. The informants are generally called correspondents.

Merits:

1. Extensive information can be had
2. It is the most cheap and economical method
3. Speedy information is possible.
4. It is useful where information is needed regularly.

Demerits:

1. The information may be biased.
2. Degree of accuracy can not be maintained.
3. Uniformity cannot be maintained.
4. Data may not be original.

4. Mailed questionnaire method:

In this method, a questionnaire consisting of a list of questions pertaining to the enquiry is prepared. There are blank spaces for answers. This questionnaire is sent to the respondents, who are expected to write the answers in the blank spaces. A covering letter is also sent along with the questionnaire. To get quick and better response, the return postage expense is borne by the investigator.

Merits:

1. Of all the methods, the mailed questionnaire method is the most economical
2. It can be widely used, when the area of investigation is large.
3. It saves money, labour, and time.
4. Error in the investigation is very small, because information is obtained directly from the respondents.

Demerits:

1. In this method, there is no direct contact between the investigator and the respondent.
2. This method is suitable only for literate people.
3. There is long delay in receiving questionnaires duly filled in.
4. People may not give the correct answer and thus one is led to false conclusion.

5. Schedules sent through enumerators:

A number of enumerators are selected and trained. They are provided with Standardized questionnaires. Each enumerator will be in charge of a certain area. The investigator goes to the informants along with the questionnaire and gets replies to the questions in the schedule and records their answers. He explains clearly the object and the purpose of the enquiry. Population census is conducted by this method.

Merits:

1. This method is very useful in extensive enquiries
2. It yields reliable and accurate results.
3. The scope of the enquiry can also be greatly enlarged.
4. Even if the respondents are illiterate, this technique can be widely used.

Demerits:

1. This is a very costly method, as the enumerators are trained and paid for.
2. This method is time-consuming, because the enumerators go personally to obtain the information
3. Personal bias of the enumerators may lead to false conclusion.
4. It is not suited to all persons due to its costliness.

Explain the methods of collecting secondary data?

Secondary data:

The various sources of secondary data can be divided into two categories.

1. Published sources 2. Unpublished sources

(1) Published Sources:

1. International agencies and bodies publish regular and occasional reports on economic and statistical matters. They are UNO, WHO, etc.,
2. Department of the Union and state governments regularly publish reports on a number of subjects. They gather additional information.
   Some of the important publications are The Reserve Bank of India Bulletin, Census of India, Agricultural Statistics of India, Indian Trade Journal, etc,.
3. Semi- Government institutions like Municipal corporation, District Board,Panchayat etc, publish reports

(2) Unpublished Sources:

There are various sources of unpublished data. They are the records maintained by various government, and private offices, the researches carried out by individual research scholars in the universities or research institutes.

How do you frame a questionnaire?

What are the steps involved for framing a questionnaire?

The questionnaire is the media of communication between the investigator and the respondents. The success of an investigation depends on the construction of the questionnaire. It requires great care, skill, wisdom, efficiency and experience. There are no hard and fast rules to be followed.

The following steps are involved for framing a questionnaire

1. The questionnaire should be brief
2. The questions should be simple to understand
3. Questions should be arranged logically
4. There must be choice (1) Simple alternative questions (2) Multiple choice questions
5. Proper words should be used in the questionnaire
6. Questions of a sensitive and personal nature should be avoided
7. Necessary instructions should be given to the informant.
8. A questionnaire should look attractive.

Classification of data

Define classification of data

The process of grouping a large number of individual facts or observations on the basis of similarity among the items is called classification.

Classification is the process of arranging things in groups or classes according to their resemblances and affinities and giving expression to the unity of attributes that may subsist amongst a diversity of individuals
- R.L. Conner

Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts - Secrist.

What are the chief characteristics of classification?

1. All the facts are classified into homogeneous groups by the process of classification
2. The classification may be according to either similarities or dissimilarities.
3. It should be flexible to accommodate adjustments.

What are the objects of Classification?

Objects of classification:

1. To condense the mass of data
2. To present the facts in a simple form
3. To bring out clearly the points of similarity and dissimilarity
4. To facilitate comparison
5. To bring out the relationship
6. To prepare data for tabulation
7. To facilitate the Statistical treatment of the data.

Types of classification

Explain in detail the various types of classification.

There are four important types of classification

1. Geographical or area wise or region wise or district wise
2. Chronological or historical i.e on the basis of time
3. Qualitative by character or by attributes
4. Quantitative or numerical or by magnitudes

1. Geographical or spatial classification

In this type of classification the data are classified according to geographical region or place. For ex, the production of paddy in different states in India, Production of wheat in different countries etc.

Country	Yield of Wheat(in kg/acre)
America	1728
China	4737
Denmark	7478
France	4739

2.Chronological classification

In this classification, the collected data are arranged according to the order of time expressed in years, months, weeks, etc, .The data generally classified in ascending order of time. For ex,

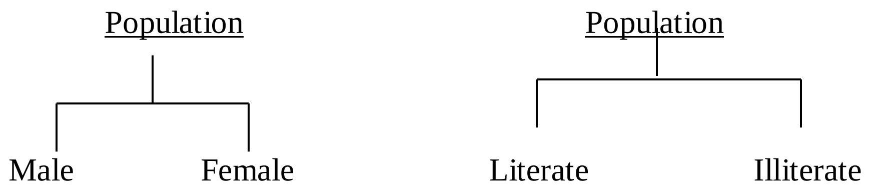
Year	:1970	1971	1972	1973	1974
Production : (in tones)	25	26	29	32	35

3.Qualitative classification

In this type of classification data are arranged on the basis of some attributes or quality like sex, intelligence, colour, literacy, religion, employment etc. Such attributes cannot be measured along with a scale.

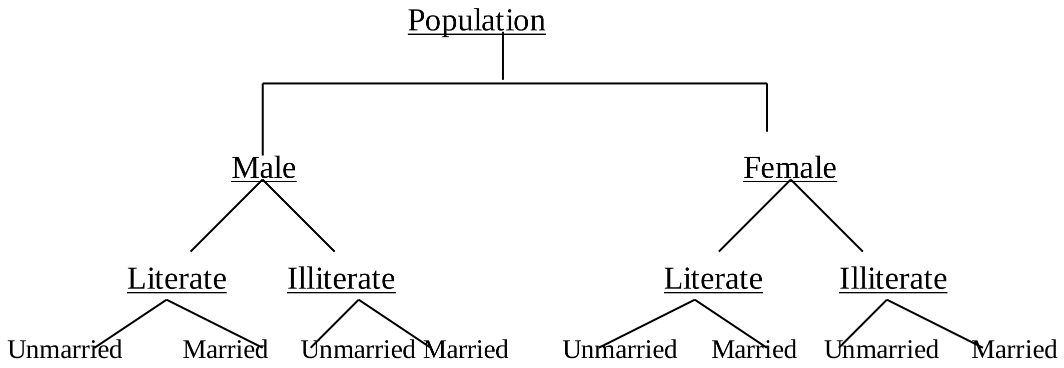
(1)Simple classification

If the data are classified into only two classes, is termed as simple classification



(2) Manifold classification:

In manifold classification, the universe is classified on the basis of more than one attribute at a time. For ex,



4.Quantitative classification

The arrangement of statistical data according to numerical measurements such as age, height, weight, amount of saving, number of members in a family come under quantitative classification. For ex, candidates who appeared in an exam can be classified according to marks obtained by them.

Marks:	0-10	10-20	20-30
No.of students:	5	7	10

Tabulation:

What do you mean by tabulation?

By tabulation we mean, a systematic presentation of numerical data in columns and rows in accordance with some salient features or characteristic. Columns are vertical arrangement and rows are horizontal arrangement. Tabulation is a process of condensing classified data in the form of a table, so that it may be more easily understood and any comparisons involved may be more readily made.

What are the objectives of tabulation?

The main objectives of tabulation are

1. To simplify complex data
2. To clarify the characteristic of data
3. To facilitate comparison
4. To detect errors and omission in the data
5. To help reference
6. To facilitate statistical processing.

Explain in detail various parts of tabulation

Parts of tabulation:

1. Table number

A table should be numbered for easy reference and identification. This number if possible, should be written in the center at the top of the table.

2. Title of the table

Each table should be given a suitable title. It must be written on the top of Table. A complete title has to answer the questions, what, where, and when in that sequence.

3. Head Note

It is a statement, given below the title and enclosed in brackets. For example The unit of measurement is written as a head note, such as “in millions” or “ in hundreds”

4.Captions or Column Headings

These are headings for the vertical columns. They must be brief and self-explanatory. They have main heading and sub-headings and must be written in small letters.

5. Stub or Row heading

These are the headings or designation for the horizontal rows. The stubs are usually wider than column headings.

6. Body of the table

The body of the table contains the numerical information in the different cells. This arrangement of data is according to the description of captions and stubs.

7. Footnote

Footnotes are given at the foot of the table for explanation of any fact or information included in the table, which needs some explanation.

8.Source Note

One should also mention the source of information from which data are taken. This may include the name of the author, volume, page, and year of publication.

What the rules to be adopted for Tabulation?

RULES FOR TABULATION:

1. The table should be simple and compact. It should not be overloaded with details.
2. The captions and stubs in the tables should be arranged in a systematic manner. It must be easy to read the important items. They are alphabetical, chronological, geographical, conventional, etc.
3. It should suit the purpose of the investigation.
4. The unit of measurements should be clearly defined and given in the tables; for ex, height in metres, weight in kilograms, etc.
5. Figures may be rounded off to avoid unnecessary details in the table. But a foot-note must be given to this effect.
6. A table should be complete and self-explanatory.
7. A table should be attractive to draw the attention of readers.
8. Abbreviations should be avoided.
9. Do not use ditto marks that may be mistaken.
10. Proper lettering will help to adjust the size of the table.

What are the differences between Classification and Tabulation?

1. Both classification and tabulation are important for statistical investigation. First the data are classified; then they are presented in tables; classification is the basis for tabulation.
2. Tabulation is a mechanical function of classification, because in tabulation classified data are placed in columns and rows.
3. Classification is a process of statistical analysis; tabulation is a process of presenting data in suitable structure.

Presentation of data:

Charts

Charting data:

What do you mean by chart?

One of the most convincing and appealing ways in which data may be presented in through charts. Evidence of this can be found in the financial pages of newspapers, journals, advertisements, etc., Pictorial presentation helps in quick understanding of the data. As the number and magnitude of figures increases, they become more confusing and their analysis tends to be more strenuous. A picture is said to be worth 10000 words. i.e through pictorial presentation data can be presented in an interesting form. Not only this, charts have greater memorizing effect as the impression created by them last much longer than those created by the figures.

A chart can take the shape of either a diagram or a graph. For the sake of clarity we will discuss them under two separate heads (1) Diagrams (2) Graphs.

Diagrams

For representing data diagrams are more commonly used than graphs. However, before discussing types of diagrams it would be worth while to consider some general rules for constructing diagrams.

What are the general rules for constructing Diagrams?

1. Title

Every diagram must be given a suitable title. The title may be given either at the top of the diagram or below it.

2. Proportion between width and height

A proper proportion between the height and width of the diagram should be maintained. If either the height or width is too short or too long in proportion, the diagram would give an ugly look.

3. Selection of appropriate scale

The scale showing the values should be in even numbers or in multiples of five or ten ex 20,50,75 or 20, 40, 60. Odd values like 1,3,5,7 should be avoided. The scale should specify the size of the unit and what it represents for ex, ‘millions’ ‘tonnes’ etc,.

4. Foot notes

In order to clarify certain points about the diagram foot notes may be given at the bottom of the diagram.

5. Index

An index illustrating different types of lines or different shades, should be given so that the reader can easily make our the meaning of the diagram.

6. Neatness and cleanliness

Diagrams should be absolutely neat and clean.

7. Simplicity

Diagrams should be as simple as possible so that the reader can understand their meaning clearly

What are the types of diagrams?

1.One dimensional diagrams Ex Bar diagrams

2. Two dimensional diagrams Ex Rectangles, Squares and circles

3. Pictograms

Explain about one dimensional or Bar diagram.

Bar diagrams are the most common type of diagrams used in practice. A bar is a thick line whose width is shown merely for attention. They are called one-dimensional because it is only the length of the bar that matters and not the width.

What are the points to be kept in mind while constructing bar diagrams?

1. The width of the bars should be uniform throughout the diagram
2. The gap between one bar and another should be uniform throughout
3. Bars may be either horizontal or vertical. Vertical bars should be preferred because they give a better look.

Types of bar diagrams:

1. Simple Bar diagram
2. Sub-divided bar diagram or component
3. Multiple bar diagram or compound
4. Percentage bar diagram.

Explain about simple bar diagram.

A simple bar diagram is used to represent only one variable. For ex, the figures of sales, production, population etc,. For various years may be shown by means of a simple bar diagram. Since the bars are of the same width and only the length varies, it becomes very easy for the reader to study the relationship. Simple bar diagrams are very popular in practice.

Explain about multiple bar diagram.

These diagrams are used to represent various parts of the total. For ex, the number of employees in various departments of a company may be represented by a subdivided bar diagram. While constructing such a diagram the various components in each bar should be kept in the same order. To distinguish between the different components, it is useful to use different shades or colours. Index or key should be given explaining these differences. Sub divided bar diagrams can be vertical as well as horizontal.

Explain about Multiple bar diagram.

In multiple bar diagram two or more sets of interrelated data are represented. The technique of drawing such a diagram is the same as that of simple bar diagrams. The only difference is that since more than one phenomenon is represented, different shades, colours, dots are used to distinguish between the bars .

Explain about Percentage bar diagram.

Percentage bar diagrams are particularly useful is Statistical work which requires the portrayal of relative changes in data. When such diagrams are prepared, the length of the bars is kept equal to 100 and segments are cut in these bars to represent the percentages of an aggregate.

Graphs:

Broadly the various graphs can be divided under the following two heads:

1. Graphs of time series or line graphs.
2. Graphs of frequency distribution.

Graphs of time series or line graphs.

The technique of graphic presentation is extremely helpful in analyzing changes at different points of time. On the X axis we generally take the time and on the Y axis the value of the variable and join the various points by straight lines. The graph so formed is known as line graph. Such graphs are most widely used in practice. They are simplest to under stand, easiest to make and most adaptable to many uses. Also several variables can be shown on the same graph and a comparison can be made.

One of the fundamental rules while constructing graphs is that the scale on the Y axis should begin from zero even if the lowest Y-figures associated with any X-period or value is far above zero. However, if this rule is strictly followed the curve would be very much pulled up towards the right ie away from the point of origin.

Graphs of frequency distribution:

A frequency distribution can be presented graphically in any of the following ways.

1. Histogram 2. Frequency polygon 3. Smoothed frequency curve 4. Cumulative frequency curve

What is meant by Histogram?

Out of several methods of presenting a frequency distribution, histogram or the column diagram ,as it is sometimes called, is the most popular and widely used in practice. The statistical meaning of histogram is that it is a graph that represents the class frequencies in a frequency distribution by vertical adjacent rectangles.

A histogram is a graphical method for presenting data, where the observations are located on a horizontal axis and frequency of those observations is depicted along the vertical axis.

While constructing histogram the variable is always taken on the X axis and the frequencies depending on it on the Y axis. Each class is then represented by a distance of the scale that is proportional to its class-interval. The distance for each rectangle on the X axis shall remain the same in the case the class-intervals are uniform throughout. The Y axis represents the frequencies of each class which constitute the height of its rectangle.

Explain about Frequency polygon.

A frequency polygon is a graph of frequency distribution. It has more than four sides. We may draw a histogram of the given data and then join by straight lines the mid-points of the upper horizontal side of each rectangle with the adjacent rectangle. The figure so formed is called frequency polygon.

Another method of constructing frequency polygon is to make the mid points of the various class-intervals and then plot the frequency corresponding to each point and to join all these points by straight lines. The figure obtained would exactly be the same as obtained by the other method. The only difference is that we do not have to construct a histogram.

Explain about Smoothed frequency curve

A smoothed frequency curve can be drawn through the various points of the polygon. The curve is drawn freehand in such a manner that the area included under the curve is approximately the same as that of the polygon. The curve should look as regular as possible and all sudden turns should be avoided.

For drawing smoothed frequency curve it is necessary to first draw the polygon and then smooth it out. The polygon can be constructed even without first constructing a histogram by plotting the frequencies at the mid-points of class intervals. The curve should begin and end at the base line, as a general rule, it may be extended to the mid-points of the class-intervals just outside the histogram.

Explain about Cumulative frequency curves or Ogives.

When frequencies are added, they are called cumulative frequencies. These frequencies are then listed in a table called a cumulative frequency table. The graph of such a distribution is called a cumulative frequency curve or an Ogive.

They are two methods of constructing ogive namely

(1) Less than method: In less than method we start with the upper limits of the classes and go on adding the frequencies. When these frequencies are plotted we get a rising curve.

(2) More than method: In more than method we start with the lower limits of the classes and from the total frequencies. We subtract the frequencies of each class. When these frequencies are plotted we get a declining curve.

MEASURES OF CENTRAL TENDENCY:

What do you mean by measures of central tendency?

Suppose the students from two or more classes appeared in the examination and we wish to compare the performance of the classes in the exam or wish to compare the performance of the same class. When making such comparisons, it is not possible to compare the full frequency distributions of marks. Therefore for such statistical analysis, we need a single representative value that describes the entire mass of data. This single representative value is called the central value or an average. This central value or an average enables us to get the information of the entire mass of data, and its value lies some where in the middle of the two extremes of the given observations. For this reason such a central value or an average is called a measure of central tendency. It is also called as the averages of first order.

Explain the characteristics of good average. (or)

What are the characteristics of good average

1.It should be rigidly (properly) defined (2) It should be easy to understand
3.It should be easy to calculate. (4) It should be based on all the items in the data
5.It should be capable of further algebraic treatment (6) It should not be affected by any single value or group of items.(7).It should be least affected by sampling fluctuations.

Write down the various definitions for Average.

1.Average is a value, which is typical or representative of a set of data- Murry R.Speigal
2. An average is a single number describing some features of a set of data- Wallis and Roberts

3.Average is an attempt to find one single figure to describe whole of figures- Clark and Sekkade

4.An average value is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is also called a measure of central value- Croxton and Cowden.

It is clear from the above definitions that an average is a single value that represents a group of values. Such a value is of great significance because it depicts the characteristics of the whole group. Since an average represents the entire data, its value lies somewhere in between the two extremes. i.e the largest and the smallest items. For this reason an average is frequently referred to as a measure of central tendency.

What are the functions of an average?

1. To facilitate quick understanding of complex data
2. To facilitate comparison
3. To know about the universe from a sample
4. To help in decision making
5. To establish mathematical relationship

Arithmetic mean:

Arithmetic average is also called as Mean. It is the most common type and widely used measure of central tendency. Arithmetic average of a series is the figure obtained by dividing the total value of the various items by their number. There are two types of arithmetic average.

1.Simple arithmetic average

Arithmetic mean is frequently referred to simply as the **Mean** and we talk about such values as mean income, mean expenditure, mean mark etc,. The simple arithmetic mean of a series is equal to the sum of variables divided by their number.

2. Weighted arithmetic average.

One of the limitations of simple arithmetic mean is that it gives equal importance to all the items of the distribution. In certain cases relative importance of the items in the distribution is not the same. Where the importance of the items varies, it is essential to allocate weights to the items. Weighted average can be defined as an average whose component items are multiplied by certain values (weights) and the aggregate of the products are divided by the total of weights.

Correcting incorrect values:

It sometimes happens that due to an oversight or mistake in copying, certain wrong items are taken while calculating mean. The problem is how to find out the correct mean. The process is very simple. From incorrect $\sum x$ deduct wrong items and add correct items and then divide the correct $\sum x$ by the number of observations. The result, so obtained, will give the value of correct mean.

What are the mathematical properties of Arithmetic mean?

Mathematical properties of Arithmetic Mean:

1.The sum of the deviations of a given set of observations from the arithmetic mean is equal to zero.

i.e. $\sum(x\text{-Mean})=0$

2.The sum of squares of deviations of a set of observations from arithmetic mean in minimum. In other words, the sum of squared deviations from mean is less than the sum of the squared deviations from any other value.

i.e. $\sum(x\text{-Mean})^2 \leq \sum(x\text{-}A)^2$

3.If every value of the variable x is increased (decreased) by a constant, the arithmetic mean of the observations so obtained also increases (or decreases) by the same constant.

4.If the values of the variable x are multiplied (or divided) by a constant, the arithmetic mean of the new observations can be obtained by multiplying (or dividing) the initial arithmetic mean by the same constant.

What are the merits and demerits of mean?

Merits:

1. It is easy to understand (2) It is easy to calculate

(3) It is based on all the observations (4) It is least affected by sampling fluctuations.

(5) It provides a good basis for comparison (6) It can be used for further analysis and algebraic treatment.

(7) The mean is a more stable measure of central tendence(Ideal average)

Demerits:

1.The mean is affected by the extreme items. (2) It cannot be located by graphical method.
(3) It can't be calculated if even a single observation in the series is missing.
(4) In a distribution with open-end classes the mean cannot be determined.
(5) It may lead to false conclusion.
(6) It is not useful for the study of qualities like intelligence, honesty and character.

What is meant by Weighted Arithmetic mean?

One of the limitations of the arithmetic mean discussed is that it gives equal importance to all the items. But there are cases where the relative importance of the different items in not the same. When this is so, we compute weighted arithmetic mean. The term weight stands for the relative importance of all the different items.

MEDIAN:

What do you mean by median?

Median is the value of item that goes to divide the series into equal parts. Median may be defined as the value of that item which divides the series into two equal parts, one half containing values greater than it and the other half containing values less than it. Therefore, the series has to be arranged in ascending or descending order, before finding the median. In other words, arranging is necessary to compute median.

As distinct from the arithmetic mean, which is calculated from the value of every item in the series, the median is what is called a positional average. The term position refers to the place of a value in a series.

What are the various definitions of Median?

According to Yau Lun Chou The median, is the value of the middle item in a series, when items are arranged according to magnitude.

The median is that value of the variable, which divides the group into two equal parts, one part comprising all the values greater, and the other, all values less than median. L.R.Conner

The median is that value which divides a series so that one half or more of the items are equal to or less than it and one half or more of the items are equal to or greater than it.- Croxton and Cowden.

Write down the mathematical properties of median.

(1)Median is an average of position
(2) The sum of the absolute deviations of the observations from the median is least.

What are the merits and demerits of median?

Merits:

1. It is easy to understand (2) It is easy to calculate
(3) It is not affected by extreme observations.
(4)The value of the median can also be located graphically.
(5) It can be calculated for distributions with open-end classes.
(6) Its value generally lies in the distribution

Demerits:

1.It is not based on all the observations (2) It is not capable of further algebraic treatment.
(3) In case on continuous series, the median is estimated, but not calculated.
(4) Median is not amenable to further algebraic manipulation
(5) It ignores the extreme items

Mode:

What do you mean by mode?

Mode is the most common item of a series. Mode is the value, which occurs the greatest number of frequency in a series. It is derived from the French word 'La Mode' meaning the fashion. Mode is the most typical value of a distribution, because it is repeated the highest number of times in the series. According to Croxton and Cowden, " The mode of a distribution is the value at the point around which the item tend to be most heavily concentrated.

Mode is defined as the value of the variable, which occurs most frequently in a distribution. Mode is also known as **Norm.** Mode is an important average in many situations specially in marketing studies. For ex, When we talk of most common income, most common wage, most common size of shoe, we have in mind the mode and not the mean or median. The chief feature of mode is that it is the size of that item which has the maximum frequency and is also affected by the frequencies of the neighboring items.

Reasons for going Grouping table:

1.If the maximum frequency is repeated (2) If the maximum frequency occurs in the very beginning or at the end of the distribution. (3) If there are irregularities in the distribution.

What are the merits and demerits of Mode?

Merits:

1. It is easy to understand (2) It is easy to calculate
(3) It can be determined for open-end distributions
(4) It can be located graphically.
(5) It is not affected by extreme items.

Demerits:

(1). It is not based on all the observations
(2) The value of mode cannot always be determined. In some cases we may have a bimodal or multimodal series .
(3) The value of mode is affected by the size of the class intervals, which is the basis of grouping the frequency.

Harmonic mean:

Harmonic mean is the reciprocal of the arithmetic average of the reciprocal of values of various items in the variable. The reciprocal of a number is that value which is obtained dividing one by the value. The reciprocal can be obtained from logarithm tables. The harmonic mean is generally used for averaging certain rates of speed, or prices, etc.

What are the uses of HM?

It is useful for computing the average rate of increase in profits of a concern or average speed at which a journey has been performed or the average price at which an article has been sold. The rate usually indicates the relation between two different types of measuring units that can be expressed reciprocally.

What are the merits and demerits of HM?

Merits:

(1).It is based on all the observations (2) It is capable of further algebraic treatment.
(3)It is not affected by sampling fluctuations.
(4) It is very useful for measuring average relative changes in certain types of rates or ratios.

Demerits:

1.It is not easy to understand
(2) It is difficult to calculate
(3) Its value cannot be computed when there are both positive and negative items in a series or when one or more items are zero.

Geometric mean:

Geometric mean is defined as the Nth root of the product of N items. If there are two items, we take the square root; if three, the cube root; and so on. The geometric mean is never larger than the arithmetic mean. If there are zeros or negative values in the series, the geometric mean cannot be used. Thus, the geometric mean is obtained by multiplying together all the values of the series and then calculating the root of their product corresponding to the number of items in the group. To solve a question to find out the geometric mean, help is taken from logarithms so as to save the time and labour. Therefore, geometric mean is the antilog of the arithmetic average of the logarithms of different items.

What are the merits and demerits of GM?

Merits: (1). It is based on all the observations
(2) It is suitable for further algebraic treatment
(3) It is very useful in dealing with ratio, rates etc.,
(4) It is not affected by the sampling fluctuations.
(5) It is useful in studying economic and social data.

Demerits: (1).It is defined only for positive values of the variable.
(2) It is difficult to understand
(3) Its computation is difficult.
(4) It has restricted application
(5) Non- mathematical persons cannot do calculations

Uses of geometric mean:

1. Geometric mean is highly useful in averaging ratios, percentages and rate of increase between two periods.
2. Geometric mean is important in the construction of index numbers.

How will you choose a suitable average for a given set of data?

Arithmetic mean may be preferred as its computation is based on all the items. But we cannot use it even if a single item is missing in the investigation. Median and mode may then be used for calculating the central value of the given observations. Thus while using averages, their inherent limitations should always be kept in mind. It is generally believed that arithmetic mean is the best average for all general purposes. It is always recommended in a situation when the data is properly spread out. i.e homogeneous in nature and do not show wide variations. Thus arithmetic mean is generally suitable in reporting the average height, average score, average income of a homogeneous group of individuals.

Median should be the choice when the object is to determine an average that would indicate its position in relation to other observations. It is also recommended for open-end distributions or when there are extreme observations in a series.

Mode is a suitable average when a quick, approximate and most typical measure of central value is desired. For ex, in business and commerce, the terms like modal output per machine, average size of collar or other ready-made garments and average expenditure of a student in a hostel refer to mode.

Harmonic mean is used in the computation of average speed or velocity and average prices when prices are quoted in terms of "units of commodity per rupee".

Geometric mean is most suited when small weight is given to large values and large weight to small values. Thus, its use is recommended for averaging rates, ratios and percentages .Due to this, it is widely used in the construction of index numbers.

UNIT-II

Measures of Dispersion (or variation):

What do you mean by measures of dispersion? Or Explain the concept of measures of dispersion.

A measure of central tendency or an average is a single representative value for a set of observations or a frequency distribution. It tells us where the center of the set of data lies but does not tell us how the set of observations is scattered around this central value. Two sets of data may have the same averages but the items in one may scatter widely around its average while in the other case, observations may be close to the average. In this way the central value or an average alone cannot describe the distribution adequately. A further description about the scatteredness is necessary to get a better description of data. The extent or degree to which data tend to spread around an average is called the dispersion or variation. Measures of dispersion help us in studying the extent to which observations are scattered around the average or central value. It is also known as the averages of Second Order.

Significance of measuring variation:

1. To determine the reliability of an average
2. To serve as a basis for the control of the variability
3. To compare two or more series with regard to their variability
4. To facilitate the use of other statistical measures.

Explain the characteristics (or properties) of good Measure of dispersion. (or)
What are the characteristics of good Measure of dispersion

1.It should be rigidly (properly) defined (2) It should be easy to understand
3.It should be easy to calculate. (4)It should be based on all the items in the data
5.It should be capable of further algebraic treatment (6) It should not be affected by any single value or group of items. (7) It should be least affected by sampling fluctuations.

Write down the various definitions for measures of dispersion.

Dispersion is the measure of the variation of the items- A.L.Bowley

Dispersion is a measure of extent to which the individual items vary. - L.R. Conner.

A measure of dispersion describe the degree of scatter shown by observations and is usually measured as an average deviation about some central value. John I Griffin

The measurement of the scatter ness of the mass of figures in a series about an average is called measure of variation or dispersion. Simpson and Kafka.

What are the two types of measures of dispersion?

(1) Absolute measure of dispersion

(2) Relative measure of dispersion

Absolute measures of dispersion are expressed in the same unit in which the observations are given.

Relative measures of dispersion are expressed as the ratio or percentage or the coefficient of the absolute measure of dispersion. Relative measures are useful for comparing variability in two or more distributions where units of measurement may be different.

Range: Range is defined as the difference between the largest value and the smallest value. Its measure depends upon the extreme items and not on all the items. It does not tell us anything about the distribution of values in the series. For purpose of comparison a relative measure of range is calculated.

What the merits and demerits of Range:

Merits

(1) It is easy to understand (2) It is easy to calculate

(3) It gives a rough but quick answer.

Demerits:

(1) It is not based on all the observations (2) It is affected by extreme items

(3) Range cannot be calculated for open-end distributions.

Uses of range:

1.Range is used in industries for the statistical quality control of the manufactured product by the construction of control chart.

2.Range is useful in studying the variations in the prices of stock, shares and other commodities that are sensitive to price changes from one period to one period.

3. The meteorological department uses the range for weather forecasts since public is interested to know the limits within the temperature is likely to vary on a particular day.

Quartile deviation:

What do you mean by Quartile deviation?

Range is based on two extreme items and it does not take into account the variation within the range. For this reason, interquartile range is defined. By eliminating the lowest 25% and the highest 25% of items in a series, Interquartile range includes the middle fifty percent of the distribution. In other words interquartile range is the difference between the third quartile and the first quartile .For comparing two or more distributions in respect of variation, the coefficient of quartile deviation is defined.

Quartile deviation gives the average amount by which the two qualities differ from the median. In asymmetrical distribution the two quartiles Q1 and Q3 are equidistant from the median and as such the difference can be taken as a measure of dispersion. When quartile deviation is very small, it describes high uniformity or small variation of the central 50% items and a high quartile deviation means that the variation among the central items is large.

What are the merits and demerits of QD?

Merits:

(1)It is simple to understand

(2)It is easy to calculate

(3) It is not affected by extreme values.

(4) It can be found out with open-end distribution.

Demerits:

(1) It ignores the first 25% of the items and the last 25% of the items .

(2) It is a positional average

(3) It is not based on all the observations.

Mean deviation or average deviation:

What does MD mean?

The range and quartile deviations are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation is measure of dispersion based on all items in a distribution. The MD is the arithmetic mean of the deviations of a series computed from any measure of central tendency that is mean, median or mode, all the deviations are taken as positive and negative signs are ignored. According to Clark and Schekade, average deviation is the average amount of scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviation. The average that is taken of the scatter is an arithmetic mean, which accounts for the fact that this measure is often called "mean deviation". The relative mean deviation of coefficient or coefficient of mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

What are the merits and demerits of MD?

Merits

(1) It is easy to understand
(2) It is easy to calculate
(3) It is based on all the observations
(4) It is not affected by extreme observations.

Demerits:

(1) In the calculation of MD, the algebraic signs of the deviations are ignored.
(2) It is not used in further calculations
(3) It is rarely used.
(4) This method may not give us very accurate results.

What do you mean by standard deviation?

Karl Pearson introduced the concept of standard deviation in 1893. It is one of the most popular and important measures of dispersion. It satisfies most of the properties of a good measure of dispersion. The SD is defined as the square root of the arithmetic mean of the squares of the deviations of the observations from the arithmetic mean. It is also known as root mean square deviation. The square of the SD is known as variance.

It may be noted that in calculating MD we ignore signs of deviations and consider their absolute values only, whereas, we square the deviations in computing SD. More so the deviations used in calculating SD are always taken from arithmetic mean only and no other central value is used for the purpose. The greater the amount of dispersion, the greater SD. On the other hand, a smaller SD means a higher degree of uniformity of the observations.

What are the algebraic properties of SD?

(1) The value of SD of a series remains unchanged if each variate value is increased or decreased by the some constant value (i.e SD is independent of change of origin).
(2) If the values of the variable x are multiplied(or divided) by a constant, the SD of the new observations can be obtained by multiplying (or dividing) the initial SD by the same constant.(i.e SD is affected by change of scale).

What are the merits and demerits of SD?

Merits

(1) It is rigidly defined
(2) Its computation is based on all the observations.
(3) Among all the measures of dispersion, it is least affected by sampling fluctuations.

Demerits:

(1) SD is comparatively difficult to calculate.
(2) It is an absolute measure of dispersion and can not be used for comparing variability of two or more distributions expressed in different units.

What do you mean by Coefficient of Variation(C.V)

It is clear that the standard deviation, as a measure of dispersion, gives us an idea about the extent to which observations are scattered around their mean. Thus, two or more distributions having the same mean can be compared directly for their variability with the help of corresponding standard deviation. Now the following two situations may arise

1. When two or more distributions having unequal means are to be compared in respect of their variability

2. When two or more distributions having observations expressed in different units of measurements are to be compared in respect of their scattered ness or variability.

For making comparisons in the above two situations, we use a relative measure of dispersion, called coefficient of variation.

Remarks:

1. Co-efficient of variation is pure number independent of the units of measurements.
2. This is useful for making comparisons between two or more distributions in respect of their variability, homogeneity, uniformity or consistency.
3. The distribution having greater C.V is considered more variable than the other, and the distribution with lesser C.V shows greater consistency, homogeneity and uniformity.

Difference between standard deviation and mean deviation

Mean deviation

1. Deviations are calculated from mean, median or mode.
2. Algebraic signs are ignored while calculating mean deviation
3. It is simple to calculate
4. It lacks mathematical properties, because algebraic signs are ignored.

Standard deviation

1. Deviations are calculated only from mean
2. Algebraic signs are taken into account
3. It is difficult to calculate
4. It is mathematically sound, because algebraic signs are taken into account.

Explain the procedure for drawing Lorenz curve?

Or What are the steps involved in drawing Lorenz curve?

A graphic method of showing dispersion is adopted by Mr Dr.Max .O. Lorenz. Lorenz curve is a device used to show the measurement of economic inequalities. The following is the method for constructing Lorenz curve.

(1) The size of item and their frequencies are to be cumulated.

(2) Percentage must be calculated for each cumulation value of the size and frequencies of items.

(3) Plot the percentage of the cumulated values of the variable against the percentage of the corresponding cumulated frequencies. Join these points with a smooth free-hand curve. This curve is known as Lorenz curve.

(4) The zero percentage on the X-axis must be joined with 100% on Y-axis. This line is called Line of equal distribution. The greater the distance between the curve and the line of equal distribution, the greater the dispersion. If the Lorenz curve is nearer to the line of equal distribution, the dispersion is smaller.

Skewness:

What are the various definitions of Skew ness?

When a series is not symmetrical it is said to be asymmetrical or skewed- Croxton and Cowden

Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution-Morris hamburg

Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median, mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness- Simpson and Kalfa

Explain the concept of Skewness or what is meant by Skewness? Or How Skewness is differed from central tendency and dispersion?

Averages determine the central point of distribution, but they give no information about the shape of the frequency curve. Measure of dispersion gives some idea of the spread of a variable about its average. Both these measures do not study whether a distribution is symmetrical or not. Skewness is a measure to study this aspect of a statistical distribution. When the items in a series are dispersed about the central value in even fashion, the frequency curve representing the distribution will be symmetrical. We can draw a graph with the help of the given frequency distribution. If the shape of the curve, or histogram, is equal on either side of the median, it is clear that the distribution is symmetrical. If we fold the curve or histogram on the ordinate at mean the two halves will coincide. It means the distribution is symmetrical. If a distribution is not symmetrical we say that it is skewed.

In a perfectly symmetrical distribution, Mean, Median and Mode coincide.

If the frequency curve has a long tail to the right, we say that it is skewed to the right.
If the frequency curve has a long tail to the left, we say that it is skewed to the left.

Tests of skew ness: Skew ness is present if,

- The values of mean, median and mode do not coincide
- When the data are plotted on a graph they do not give the normal bell-shaped form
- The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
- Quartiles are not equidistant from the median
- Frequencies are not equally distributed at points of equal deviation from the mode.

When skew ness is absent

- The values of mean, median and mode coincide
- Data when plotted on a graph give the normal bell-shaped form
- The sum of the positive deviations from the median is equal to the sum of the negative deviations.
- Quartiles are equidistant from the median
- Frequencies are equally distributed at points of equal deviation from the mode.

UNIT-III

CORRELATION

What is meant by correlation?

Or Explain about correlation.

This is one of the method of studying the relationship between the variables. In the study of two variables the change in the value of one variable produces a change in the value of other variable. In that case we say that the variables are correlated or there is a correlation between the two variables. These two variables may have a positive correlation, a negative correlation or they may be uncorrelated.

The association of any two variates is known as correlation. Correlation is the numerical measurement showing the degree of correlation between two variables. One variable may be called subject(independent) and the other relative(dependent) variable. Independent variables are measured in terms of dependent variables.

What are the important definitions of correlation?

According to Ya Lun Chou ,correlation analysis attempts to determine the degree of relationship between variables.

According to W.I.King correlation means that between two series or group of data there exists some casual connection.

According to A.M.Tuttle correlation is an analysis of the co variation between two or more variables

Explain the types of correlation with examples.

The important types of correlation are

1.positive and negative 2.simple and multiple 3.partial and total 4.linear and non-linear.

1.POSITIVE AND NEGATIVE CORRELATION:

Two variables are said to be positively correlated correlated if for an increase in the value of one variable there is also an increase in the value of the other variable or for a decrease in the value of one variable there is also a decrease in the value of the other variable, that is the two variables change in the same direction.

Ex:Years of experience and salary of employees in a company ,Height and weight,rainfall and yield of crops.

Two variables are said to be negatively correlated if for an increase in the value of one variable there is a decrease in the value of the other variable, that is the two variables change in opposite direction.

Ex:when the price increases, the demand for the commodity decreases and when the price decreases the demand increases.

2.SIMPLE AND MULTIPLE CORRELATION:

The distinction between simple,partial,and multiple correlation is based upon the number of variables studied. When only two variables are studied, it is a problem of simple correlation. When three or more variables are studied, it is a problem of either multiple and partial correlation. In multiple correlation three or more variables are studied simultaneously.

For ex,when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, is a problem of multiple correlation.

3.PARTIAL AND TOTAL CORRELATION:

The study of two variables excluding some other variables is called partial correlation. For ex,we study price and demand eliminating the supply side. In total correlation ,all the facts are taken into account.

4.LINEAR AND NON-LINEAR:

If the ratio of change between two variables is same, then there will be linear correlation between them. Consider the following

X:	1	2	3	4
Y:	3	6	9	12

The ratio of change between the variables is the same. If we plot these points on a graph, we get a straight line. In a non-linear(curvi linear) correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variables. The graph of non-linear relationship will form a curve. However, since techniques of analysis for measuring non-linear correlation are more complicated than those for linear correlation, we generally make an assumption that the relationship between the variables is of the linear type

NO CORRELATION:

Two variables are said to be uncorrelated, if the change in the value of one variable has no connection with the change in the value of the other variable.For ex, weight of a person and the colour of his hair ,the height of a person and the colour of his hair.

What are the uses of correlation?:

1.Correlation is very useful to economists to study the relationship between variables like price and quantity demanded. The businessmen, it helps to estimate costs,sales,price and other related variables.

2.Correlation analysis helps in measuring the degree of relationship between the variables,like supply and demand,price and supply,income and expenditure etc,.

3.The relation between variables can be verified and tested for significance, with the help of the correlation analysis.

4.We can compare the relationship between variables which are expressed in different units.

What are the properties of correlation coefficient?

1.The correlation coefficient is unaffected by change of origin and change of scale. By change of origin ,we mean that a constant is subtracted from all the values of X and Y series.By change of scale,we mean that all the values of X and Y series are multiplied or divided by some constant.

2.Coefficient of correlation lies between -1 and +1

i.e $-1<=r<=+1$

When r is +ve ,the variables X and Y increase or decrease together.

When r is -ve ,the variables X and Y move in the opposite direction.

When r=+1,implies that there is a perfect +ve correlation between X and Y

When r=-1,implies that there is a perfect -ve correlation between X and Y

When r=0,the two variables are uncorrelated.

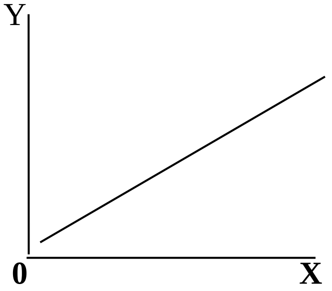
Methods of studying Correlation:

Graphic method: Explain about Simple Graph or Correlogram

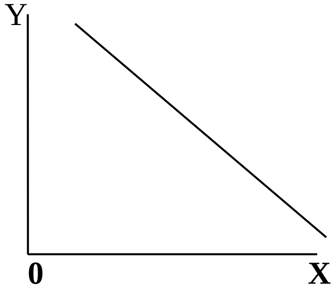
The values of the two variables are plotted on a graph paper. We get two curves, one for X variables and another for Y variables. These two curves reveal the direction and closeness of the two curves and also reveal whether or not the variables are related. If both the curves move in the same direction, ie parallel to each other, either upward or downward, correlation is said to be positive. On the other hand, if they move in opposite directions, then the correlation is said to be negative.

Explain in detail about Scatter diagram method:

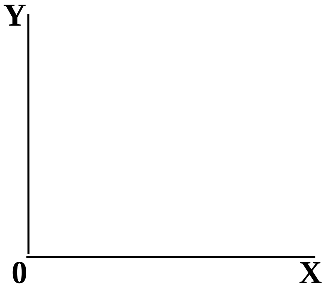
This is the simplest method of finding out whether there is any relationship present between two variables by plotting the values on chart, known as scatter diagram. In this method, the given data are plotted on a graph paper in the form of dots. X variables are plotted on the horizontal axis and Y variables on the vertical axis. Thus we have the dots and we can know the scatter of various points.' This will show the type of correlation.



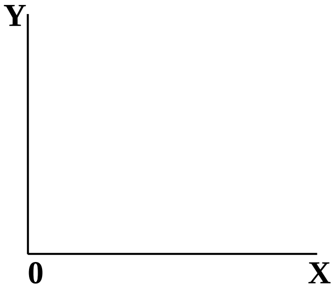
Perfect positive Correlation
(r = +1)



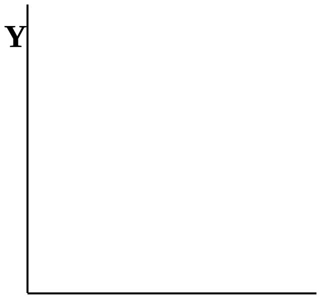
Perfect Negative Correlation
(r = -1)



High degree positive Correlation



High degree Negative Correlation



No correlation

If the plotted points form a straight line running from the lower left-hand corner to the upper right hand corner, then there is a perfect positive correlation. (i.e r=+1) . On the other hand, if the points are in a straight line, having a falling trend from the upper left-hand corner to the right-hand corner, it reveals that there is a perfect negative or inverse correlation (i.e r=-1)

If the plotted points fall in a narrow band, and the points are rising from lower left-hand corner to the upper right-hand corner, there will be a high degree or positive correlation between the variables. If the plotted points fall in a narrow band from the upper left-hand corner to the lower right-hand corner, there will be a high degree or negative correlation. If the plotted points lie scatter all over the diagram, there is no correlation between the two variables.

What are the merits and demerits of scatter diagram?

Merits

1) Scatter diagram is a simple method of finding out the nature of correlation between two variables.
2) It is easy to understand.
3) We can get a rough idea at a glance whether it is a positive or negative correlation.
4) It is a first step in finding out the relationship between two variables.

Demerits

By this method we cannot get the exact degree or correlation between two variables. It gives only a rough idea.

SPEARMAN'S RANK CORRELATION COEFFICIENT:

The Karl Pearson's formula for calculating r is developed on the assumption that the values of the variables are exactly measurable.In some situations it may not be possible to give precise values for the variables.In such cases we can use another measure of correlation coefficient called rank correlation coefficient. We rank the observations in ascending or descending order using the numbers 1,2,3....n and measure the degree of relationship between the ranks instead of actual numerical values.

REGRESSION ANALYSIS:

Regression is the study of the relationship between the variables.In regression analysis there are two types of variable. One is dependent variable and the other independent variable**.**

What are the various Definitions for regression analysis?

According to Blair,Regression is the measure of the average relationship between two or more variables in terms of the original units of data.

According to Ya-Lun Chou, Regression analysis attempts to estabilish the nature of the relationship between variables that is to study the functional relationship between the variable and thereby provide a mechanism for prediction or forecasting .

What are the two Regression Equations?

Regression analysis helps to estimate one variable or the dependent variable from the other variable or the independent variable. In other words, we can estimate the value of one variable, provided that the value of the other variable is given.

When the two variables (x and y) are plotted on a scatter diagram the two "lines of best fit" can be drawn which pass between the plotted points. These lines are called the Regression lines and the two equations based on these two lines are called the Regression Equations.

If Y is dependent variable and X is independent variable, the linear relationship suggested between the variables is called the regression equation of Y on X .Here the parameters are determined using the principle of least squares. This regression equation is used to estimate the value of Y corresponding to a known value of X.

If X is dependent variable and Y is independent variable, the linear relationship suggested between the variables is called the regression equation of X on Y .This is used to estimate the value of X corresponding to a known value of Y.

What are the properties of regression coefficients?

1.The coefficient of correlation is the geometric mean of the regression coefficients.
2.If one of the regression coefficient is greater than unity, then the other is bound to beless than unity.
3. Arithmetic mean of the regression coefficient is greater than the correlation coefficient ,provided r>0
4.Regression coefficients are independent of change of origin but not of scale.
5.The signs of both the regression coefficients are the same… i.e Either both will be +ve or both will be –ve.
6. The sign of the correlation coefficient is the same as that of the 2 regression coefficients. Thus r will be +ve ,if b_{yx} and bxy are +ve. Similarly r will be –ve if bxy and byx are –ve.
7.Both the regression coefficients cannot simultaneously exceed one.

Explain the uses of regression analysis:

1. It is used to estimate the relationship between two economic variables like income and expenditure.
2. It is very useful for prediction purposes.
3. In business, it is very helpful to study business prediction’s Cost if production is affected by the scale of production.
4. It is used to predict demand curves, supply curves, production
Function, cost function.
5.With the help of regression coefficients, we can calculate the Correlation coefficient.

DIFFERENCE BETWEEN CORRELATION AND REGRESSION:

CORRELATION

1.Correlation coefficient is independent of change of scale and origin.
2.If the correlation coefficient is +ve, then the two variables are +vely correlated.
3.There may be non-sense correlation between two variables.
4.The correlation coefficient between X and Y rxy and correlation
Coefficient between Y and X, ryx are symmetric. i.e., rxy=ryx
5.It is not very useful for further mathematical treatment.

REGRESSION

1.Regression coefficients are independent of change of origin but not of scale.
2.The regression coefficient explains that the decrease in one variable is associated with the increase in the other variable.
3.In regression there is no such non-sense regression.

4.In regression analysis the regression coefficients bxy and byx are not symmetric.

5.It is widely used for further mathematical treatment.

UNIT-IV

INDEX NUMBERS:

An index number is a measure that shares a relative comparison between groups of related items. An index number is nothing but a percentage figure that expresses the relationship between two numbers with one of the number used as the base. Our aim is to study the change in prices of a number of commodities. In this case we want to combine the price relatives of many commodities into a single figure for each a year. For such a figure we use the term Index number.

What are the various definitions for index numbers?

DEFINITIONS:

Index numbers are devices for measuring differences in the magnitude of a group of related variables.
-Croxton & Cowden.

An index number is a statistical measure designed to show changes in a variable or group related variables with respect to time, geographical location or other characteristics. -Spiegal

An index number is a percentage relative that compares economic measures in a given period with those same measures at a fixed period in the past.-Clark and Schkade.

What are the uses of index numbers?

1. Index numbers are widely used in connection with decision making and analysis in business.
2. They are very useful in measuring relative changes.
3. Index numbers reduce the changes of price level into more useful and understandable form.
4. Various index numbers computed for different purposes say employment, transport, industry etc.
5. The stability of prices or inflating or deflating conditions can be observed with the help of indices.
6. In the field of economy,the wage adjustments are done with the study of consumer's price index numbers.
7. Cost of living index numbers are used fixing the dearness allowance to the employees to enable them to meet the increased cost of living.

What are the problems involved in the construction of index numbers:

The following are some of the problems to be considered in the construction of index nos

1. PURPOSE OF INDEX NUMBERS:

An index number constructed for one purpose in general cannot be used for other purposes. For ex,for constructing the whole sale price index number, the retail prices are not necessary. Hence the purpose of index number to be constructed should be well-defined before the collection of data.

2. SELECTION OF COMMODITIES:

It is not possible to include all the commodities in the construction of index numbers. After selection of the commodities to be included in the construction of index number sampling procedure can be adopted to determine which specific prices will be included. For ex, if we study the change in production of cloth, then we may include the production of mill cloth, power loom cloth, handloom cloth, silk etc, and there is no problem.

3. CHOICE OF THE BASE:

The base period of an index number is very importance as it is used for the construction of index number. Every index number must have a base. When selecting a base period ,the year must be recent and normal. The normal year is one. If abnormal years are considered ,then the index number will be a misleading one.

4.DETERMINATION OF WEIGHTS:

Usually two types of weights are used in attaching weights. They are 1.Quantity weight 2.Value weight. Quantity weights for commodities are useful when importance is attached to the number of units of the commodities used. Value weights are useful when importance is attached to the expenditure incurred on them.

5. SELECTION OF AVERAGES:

In practice the arithmetic mean is used , because it is easy for computation. Geometric mean and Harmonic mean are difficult to calculate. But GM is preferred because of the following characteristics.

1. GM is the best measure 2.It gives less weight to bigger items and more weight to smaller items.

Explain about Methods of constructing index numbers:

They can be grouped under two heads

1. Un weighted index numbers
2. Weighted index numbers

In the un weighted indices weights are not expressly assigned where as in the weighted indices weights are assigned to the various items. Each of these types may further be divided under two heads

1. Simple aggregate
2. Simple average of price relatives

1. Simple aggregate

This is the simplest method of constructing index numbers. When this method is used to construct a price index, the total of current year prices for the various commodities in question is divided by the total of base year prices and the quotient is multiplied by 100.

2. Simple average of price relatives

When this method is used to construct a price index, price relatives are obtained for the various items included in the index and then an average of these relatives is obtained using any one of the measures of central tendency, i.e arithmetic mean, median, mode, geometric mean or harmonic mean. When arithmetic mean is used for averaging the relatives.

Weighted index numbers

Weighted index numbers are of two types

1. Weighted aggregative index number
2. Weighted average of price relative

Weighted aggregate index numbers

These index numbers are of the simple aggregative type with the fundamental difference that weights are assigned to the various items included in the index. There are various methods of assigning weights and consequently a large number of formulas for constructing index numbers.

Weighted average of price relative

In the weighted aggregative methods price relatives were not computed. However, like un weighted relatives method it is also possible to compute weighted average or relatives. For purposes of averaging we may use either the arithmetic mean or the geometric mean.

EXPLAIN THE VARIOUS TESTS FOR INDEX NUMBERS:

1. TIME REVERSAL TEST:

Time Reversal test is a test to determine whether a given method will work both ways in time, forward and backward. In the words of Fishers “The test is that the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base. In other words when the data for any two years are treated by the same method, but with the bases reversed, the two index numbers secured should be reciprocals for each other so that their product is unity. Symbolically, the following relation should be satisfied

$$\mathbf{P_{01} \times P_{10} = 1}$$

Where $\mathbf{P_{01}}$ is the index for time 1 on time 0 as base and $\mathbf{P_{10}}$ is the index for time 0 on time 1 as base. If the product in not unity, there is said to be a time bias in the method.

2. FACTOR REVERSAL TEST:

Another test suggested by Fisher is known a factor reversal test. It holds that the product of price index and the quantity index should be equal to the corresponding value index. In the words of Fisher Just as each formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results. i.e the two results multiplied together should give the true value ratio. In other words the test is that the change in price multiplied by the change in quantity should be equal to the total change in value. The total value of given commodity in a given year is the product of the quantity and the price per unit. The ratio of the total value in a one year to the total value in the preceding year is $\mathbf{\sum(p_1 q_1) / \sum (p_0 q_0)}$

EXPLAIN THE CONCEPT OF CHAIN AND FIXED BASE INDEX NUMBERS:

In the fixed base method, the base remains constant throughout. i.e the relative for all the year is based on the prices of that single year. On the other hand ,in the chain base method, the relative for each year is found out from the prices of the immediately preceding year. Thus the base changes from year to year. The indices which we find out by this method are called link relative index numbers or link relative. These link relatives are linked together

DIFFERENCE BETWEEN CBI AND FBI

CBI	FBI
1. The base year changes	1.The base year does not change.
2.Here the LR method is used	2.No such LR method is used
3.Introduction and deletion of items are easy to calculate without recalculation of the entire series	3.Any change in the commodities will involve the entire index number to be recast.
4.It is difficult to understood	4.It is simple to understand.
5.It cannot be computed if data for one year are missing.	5.There is no such problem.

WHAT DO YOU MEAN BY COST OF LIVING INDEX NUMBERS?

Consumer price index

Consumer price index is also called as the cost of living index. Statisticians recommend that the terms cost of living index or price of living index or cost of living price index or consumer price index can be used in the approximate place. In different countries, cost of living index, consumer price index and retail price index are used.

What are the uses of consumer price index?

1. This is very useful in wage negotiations and wage contracts and allowance adjustment in many countries.
2. Govt. can make use of these indices for wage policy, price policy, taxation, general economic policies and rent control.
3. Changes in the purchasing power of money and real income can be measured.
4. We can analyze the market price for particular kinds of goods and services by this index.

What are the limitations of index numbers?

1.There may be error in each stage of the construction of the index number, namely, selection of commodities, selection of the base period, selection of weight , etc

2. Index numbers may not represent the exact change in price level, because they are based on sample data.

3. Tastes, habits and customs of people change in course of time and may make the weighting not suitable for the present data.

4. In each index there is and index error, because there is no formula for measuring the price change. So there is the formula error. Hence it will not be a representative one.

5. By selecting a suitable year as the base year, selfish persons may get their desired results.

UNIT-V

ANALYSIS OF TIME SERIES:

Meaning:

An arrangement of statistical data in accordance with time of occurrence or in a chronological order is called a time series. The numerical data, which we get at different points of time-the set of observations, is known as time series. In time series analysis, current data in a series may be compared with past data in the same series.

What are the various definitions of time series?

A time series is a set of observations arranged in chronological order-Morris Ham- burg.

When quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series-Wesses & Wellet.

A time series consists of data arranged chronologically-Croxton & cowden.

Explain about the Components of time series

```
            ---------------------------------------
            |                                 |
        Long term                      Short-term
      |--------------|               |-----------------|
     secular     cyclical        seasonal     irregular(erratic)
```

1.Secular Trend:

A time series data may show upward trend or downward trend for a period of years,for ex, population increases over a period of years,prices increases over a perios of years.These are examples of upward trend.The sale of a commodity may decrease over a period of years because of better products coming to market. This is an ex for downward trend.The increase and decrease in the movements of a time series is called a secular trend.

2.Seasonal variation:

Seasonal variations are short-term fluctuations is a time series which occur periodically in a year.This continues to repeat year after year. Series of monthly and quarterly data are used to examine these seasonal variations. In general the period of seasonal variation refers to a year.But it can also take place in month or a week or a day.

Ex: Ice creams are sold during summer season

Woolen clothes are sold during winter season.

3.Cyclical variation:

Cyclical variations are recurrent upward or downward movements in a time series, but the period of cycle is greater than a year. The ups and down in business activities are the effects of cyclical variation.

4.Irregular Variations:

Irregular variations are fluctuations in time series that are short in duration. Irregular variations are caused by war,flood,strike ,revolution etc.,This variation is usually a short term one ,but it will affect all the components of time series.

What are the uses of time series analysis:

1. It helps in understanding past bahaviour and it will help in estimating the future behaviour.
2. With the help of time series we can prepare plans for future.
3. Comparison between data of one period with that of another period is possible.
4. It is very essential in business and economics and it helps in forecasting.
5. Seasonal,cyclical,secular trend of data is useful not only to economists but also to the businessman.

Explain about the methods of measuring trend.

Measures of trend:

The following are the four methods which can be used for determining the trend

1. Free-hand or Graphic method
2. Semi-average method
3. Moving average method
4. Method of least squares

1. Free-hand or Graphic method

This is the easiest, simplest and the most flexible method of estimating secular trend. In this method we must plot the original data on the graph. Draw a smooth curve carefully which will show the direction of the trend. In this the time is shown on the horizontal axis and the value of the variables is shown on the vertical axis. To the proper trend line, we must note some points, while fitting a trend line by the free-hand method.

1. The curve should be smooth.

2. Approximately there must be equal number of points above and below the curve.
3. The total deviations of the data above the trend line must be the same as the vertical deviations below the line.

Merits:

1 .It is the simplest, easiest and quickest method

2. It saves time and labor

3. It is adaptable and flexible.

4. Experienced statisticians can draw a free-hand line more accurately.

5. It will help to understand the character of time series, and we can use the appropriate mathematical trend.

Demerits:

1. It is subject to personal bias.

2. The results depend upon the judgment of the person who draws the line. There may be different curves for different persons.

3. It seems to be very simple but it requires more time for a careful job.

4. If it is not drawn by experienced persons, then it is dangerous to use for forecasting purposes.

5. It does not help us to measure trend.

2. Semi-Average method.

In this method, the original data is divided into two equal parts and averages are calculated for both the parts. These averages are called semi-averages .For example we can divide the 10 years into two equal parts. If period is odd number of years the value of the middle year is omitted. We can draw the line by a straight line by joining the two points of average. By extending the line downward or upward, we can get the intermediate values or we can predict the future values.

Merits:

1 .It is simple and easier to understand than the moving average method and least square method.

2. As the line can be extended both ways, we can get the intermediate values and predict the future values.

Demerits:

1. Under this method, is has an assumption of linear trend whether such a relationship exists or not.
2. It is affected by the limitation of arithmetic mean.
3. This method is not enough for forecasting the future trend or for removing trend from original data.

3. Moving average method.

Explain Moving Average Method

It is a simple method of indicating the trend of the given values over a period of time. Moving averages are usually calculated for 3years,4years,5 years etc,.In calculating 3 yr moving average of a time series data the average of the first 3yrs is taken to represent the trend value for the middle of these 3yrs .Then leaving the first year value ,the average of the next 3 yrs is taken as trend value of the middle year of these 3yrs. For a 3yr moving average the trend values for the first year and the last year are not known.

In similar way we can determine the 5yr moving average. Here the trend values for the first two years and the last 2yrs will be missing. This is the procedure of finding the moving average of odd period.

To find the moving average of period 4 yrs we find the average of the first 4yrs and this represents the trend value for the middle of the 2nd and 3rd year leaving the first,find the average of the next 4yrs and so on.

We once again take the average of first two trend values, the average of the next two leaving the first and so on in order that the years representing the trend value coincide with the years in the given data.This is called 4yr centred moving average . Here the trend value for the first 2 yrs and the last 2yrs will be missing.

Merits:

1. It is simple and easy to understand. It is easier to adopt when compared to the method of least squares.
2. It is more flexible than other methods. It is elastic in the sense; items can be increased or decreased without affecting the moving average.
3. It is not only used for the measurement of trend, but also for the measurement of seasonal, cyclical and irregular fluctuations.
4. The period of moving average is determined by the data and not by the personal judgment of the investigator. So it is away from personal bias.

Demerits:

1. In this method we cannot get the trend values for all the given observations. In finding trend value we leave the first and the last year in three-yearly moving average; and we leave the first two and the last two years in five yearly moving average.
2. There is no rule regarding the choice of the number of the moving average, and so the statistician has to use his own judgment.
3. In most of the economic and business time series the trend is a non-linear one; then the moving average lies below and above the curve of the actual data.

4.Method of Least squares:

By the method of least squares, a straight line trend can be fitted, to the given time series data. It is a mathematical as well as an analytical method. The trend line is called the line of best fit. The sum of deviations of the actual values of Y and the trend value (Yc) is zero and the sum of squares of deviations of the actual value and the trend value is the least.

The trend line is fitted to the data in such a manner that the following two conditions are satisfied

(1) ((Y-Yc)= 0

i.e the sum of deviations of the actual values of Y and the computed values of Y is zero.

(2) ((Y-Yc)2 is least.

i.e the sum of the squares of the deviations of the actual and computed values is least from the line. That is why this method is called the method of least squares. The line obtained by this method is known as the line of best fit. The method of least squares can be use either to fit a straight line trend or a parabolic trend.

The straight line trend is represented by the equation

Yc= a + b X

Where Yc denotes the trend(Computed) values to distinguish them from the actual Y values, "a" is the intercept or the value of the Y variable when X=0 , "b" represents slope of the linear or the amount or change in Y variable that is associated with a change of one unit in X variable. The X variable in time series analysis represents time.

In order to determine the value of the constants a and b the following two normal equations are to be tested

$$\sum Y = n\, a + b \sum (X)$$

$$\sum XY = a \sum (X) + b\,(X^2)$$

Where "n" represents number of years (months or any other time period) for which data are given

Explain about measurement of seasonal variations.

Most of the phenomenon in economics and business show seasonal patterns. When data are expressed annually there is no seasonal variation. However, monthly or quarterly data frequently exhibit strong seasonal movements and considerable interest attaches to devising a pattern of average seasonal variation. For ex, if we observe the sales of a bookseller, we find that of the quarter July-September (when most of the students purchase books), sales are maximum. If we know how much the sales of the quarter are usually above or below the previous quarter for seasonal reasons, we shall be able to answer a very basic question, namely was this due to an underlying upward tendency or simply because this quarter is usually seasonally higher than the previous quarter?

In order to analyze seasonal variation, it is necessary to assume that the seasonal pattern is superimposed on a series of values and independent of these in the sense that the same pattern is superimposed irrespective of the level of the series i.e June quarter always contributes so much more or so much less of the series.

For monthly data , a seasonal index consists of 12 numbers, one for each month of a year or number of years, that has taken place typically in each month. A specific index refers to the seasonal changes during a particular year. Seasonal indices are given as percentage or their average.

END OF UNIT V