Business statistics

Unit-1

Statistics: Definition, Origin and Growth, Functions, Applications and Limitations- Classification of data: Types of classifications.

Origin and Growth of Statistics:

The origin of statistics can be traced back to the primitive man, who put notches on trees to keep an account of his belongings. During 5000 BCE, kings used to carry out census of populations and resources of the state. Kings of olden days made their crucial decisions on wars, based on statistics of infantry, and elephantary units of their own and that of their enemies. Later it enhanced its scope in their kingdoms' tax management and administrative domains. Thus, the word 'Statistics' has its root either to Latin word 'Status' or Italian word 'Statista' or German word 'Statistik' each of which means a **'political state'**. The word 'Statistics' was primarily associated with the presentation of facts and figures pertaining to demographic, social and political situations prevailing in a state/government. It is regarded as the "Science of Statecraft". It is the by-product of States administrative activities. It has been the traditional function of Government to keep record of the population, births, deaths, taxes, crop yields and many other type of activities. Its evolution over time formed the basis for most of the science and art disciplines. Statistics is used in the developmental phases of both theoretical and applied areas, encompassing the field of Industry, Agriculture, Medicine, Sports and Business analytics.

In olden days statistics was used for political- war purpose. Later, it was extended to taxation purposes. This is evident from Kautilya's Arthasastra (324 – 300 BCE). Akbar`s finance minister Raja Thodarmall collected information regarding agricultural land holdings. During the seventeenth century, statistics entered in vital statistics, which is the basis for the modern day Actuarial Science. Gauss introduced the theory of errors in physical sciences at the end of eighteenth century.

According to a Greek historian, in 1400 B.C. a census of all the lands in the Egypt were taken. Similar reports on the ancient Chinese, Greeks and Romans are also available. People and land were the earliest objects of statistical enquiry.

Statistics came from an Italian Word “Statista” means “Statesman”, German word “Statistik” means Political State.

It was first used by a Professor Gottfried Achenwall (1719-1772), a Professor in Malborough in 1749 to refer to the subject matter as a whole. He defined as a “political science of several countries”.

The word Statistics appeared for the first time in the famous book “Elements of Universal Erudition” by Baran J.F. Von Bielfeld. It is translated by W.Hooper.

Statistics is concerned with scientific method for collecting, organizing, summarizing, presenting, analyzing and interpreting of data. The word statistics is normally referred either as numerical facts or methods.

Statistics is used in two different forms-**singular and plural**. In plural form it refers to the numerical figures obtained by measurement or counting in a systematic manner with a definite purpose such as number of accidents in a busy road of a city in a day, number of people died due to a chronic disease during a month in a state and so on. In its singular form, it refers to statistical theories and methods of collecting, presenting, analyzing and interpreting numerical figures.

Sources of Statistics Origin:

1. Government records
2. Mathematics

1. Government records:

- In ancient Eggypt, police prepared registration lists of all the heads of the families.
- In ancient Judea, census of population was taken.
- In Roman- census was recorded about military strength, taxation, births and deaths.

  As statistics is used for Government Purpose it is called as “the Science of kings” or “the Science of Statescraft”.

  William Petty was the author of “Essays on Political Arthematick”. He regarded Statistics as “Political arthemetic”.

2. **Mathematics:**
    - Statistics is a branch of Applied Mathematics. In 17th century the attention of Mathematicians like Bernoulli, Galileo, Laplace and Karl Gauss has moved towards Statistics. They developed Probability theories like winning and losing in gamble.
    - Statisticians were engaged in calculating the risk associate with a particular decision.
    - Abraham DeMoivre (1667-1754) –discovered Normal Statistical Theory.
    - Jacques Quettlet (1796-1874) – discovered the constancy of great numbers which is the basis for sampling.
    - Francis Galton(1822-1911)- regression
    - Karl pearson(1857-1936) – Chi-Square test
    - Ronald Fisher (1890-1962) – field experimental designs.

    These are the real giants in the development of the theory of statistics.

Growth of Statistics:

Though the importance of statistics was strongly felt, its tremendous growth was in the twentieth century. During this period, lot of new theories, applications in various disciplines were introduced. With the contribution of renowned statisticians several theories and methods were introduced, naming a few are Probability Theory, Sampling Theory, Statistical Inference, Design of Experiments, Correlation and Regression Methods, Time Series and Forecasting Techniques.

In early 1900s, statistics and statisticians were not given much importance but over the years due to advancement of technology it had its wider scope and gained attention in all fields of science and management. We also tend to think statistician as a small profession but a steady growth in the last century is impressive. It is pertinent to note that the continued growth of statistics is closely associated with information technology. As a result several new inter- disciplines have emerged. They are Data Mining, Data Warehousing, Geographic Information System, Artificial Intelligence etc. Now-a-days, statistics can be applied in hardcore technological spheres such as Bioinformatics, Signal processing, Telecommunications, Engineering, Medicine, Crimes, Ecology, etc.

Today's business managers need to learn how analytics can help them make better decisions that can generate better business outcomes. They need to have an understanding of the statistical concepts that can help analyze and simplify the flood of data around them. They should be able to leverage analytical techniques like decision trees, regression analysis, clustering and association to improve business processes.

Factors responsible for development of Statistics in Modern times:

1. **Increased demand for statistics:**
   - In the present century considerable development has taken place in the field of Business and commerce, Governmental activities and Science. Statistics helps in formulating suitable policies.
   - Due increase size of the Business- Statistics helps in resolving the complex problems.
   - Functions of Government are enlarged (Law and Order)- Statistics helps in proper maintenance of Law and Order.
   - Development of Science and technology- Statistics helps in doing research.

2. **Decreased cost of Statistics:**
   - Time and cost of collecting data are the important factors in the use of Statistics. With Calculators, electronic machines e.t.c the cost of analyzing the data has come down. This has led to the use of statistics in solving various problems.
   - With the development of statistics the cost of collecting and analyzing the data has come down.
   - Many have contributed to the science of statistics.

Inspite of the developments, the list of unsolved statistical problems are long and the statistical research today is more vigorous than ever before.

Definitions:

Statistics is the science of numbers. The English word Statistics originates from the Italian word “Statista” or the Latin word “Status”. Statistics is about creating information about any number-based information by searching and recording it. In other words, statistics is the scientific method of collecting, analyzing, and interpreting numerical data.

Croxton and Cowden, "Statistics may be defined as **the collection, presentation, analysis, and interpretation of numerical data**."

Edward N. Dubois defines, "Statistics is a body of methods for obtaining and analyzing numerical data in order to make better decisions in an uncertain world."

According to Bowleg, "Statistics may rightly be called the science of averages."

Prof. Boddington defines, "The science of estimate and probabilities."

According to Prof. Horace Secrist, "Statistics is the aggregate of facts affected to a marked extent by the multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other."

American Heritage Dictionary defines statistics as: "The mathematics of the Collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling."

A.L. Bowley defines, "Statistics are numerical statements of facts in any department of inquiry placed in relation to each other."

Wallis and Roberts define, "Statistics is a body of methods for making wise decisions in the face of uncertainty."

According to Prof. Ya Lun Chaou, " Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risk."

According to W.I. King, "The science of statistics is the method of judging collection, natural or social phenomena from the results obtained from the analysis or enumeration or collection of estimates."

The Merriam-Webster's Dictionary definition is: "A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."

Features of Statistics:

1. A numerical expression is required in statistics.
2. Statistics is the sum of information.
3. The search for statistics must be related to a specific field.
4. Statistical information is influenced by multiple factors.
5. Statistical information needs to be collected in a well-organized manner.
6. The statistics should be comparable and homogeneous.
7. There is a need to maintain logical and quantitative accuracy in statistical estimates.

Functions of Statistics:

1. To Present Facts in Definite Form:

We can represent the things in their true form with the help of figures. Without a statistical study, our ideas would be vague and indefinite.

The facts are to be given in a definite form. If the results are given in numbers, then they are more convincing than if the results are expressed on the basis of quality.

The statements like, there is lot of unemployment in India or population is increasing at a faster rate are not in the definite form. The statements should be in definite form like the population in 2004 would be 15% more as compared to 1990.

2. Precision to the Facts:

The statistics are presented in a definite form so they also help in condensing the data into important figures. So statistical methods present meaningful information. In other words statistics helps in simplifying complex data to simple-to make them understandable.

The data may be presented in the form of a graph, diagram or through an average, or coefficients etc. For example, we cannot know the price position from individual prices of all good, but we can know it, if we get the index of general level of prices.

3. Comparisons:

After simplifying the data, it can be correlated as well as compared. The relationship between the two groups is best represented by certain mathematical quantities like average or coefficients etc. Comparison is one of the main functions of statistics as the absolute figures convey a very less meaning.

4. Formulation and Testing of Hypothesis:

These statistical methods help us in formulating and testing the hypothesis or a new theory. With the help of statistical techniques, we can know the effect of imposing tax on the exports of tea on the consumption of tea in other countries. The other example could be to study whether credit squeeze is effective in checking inflation or not.

5. Forecasting:

Statistics is not only concerned with the above functions, but it also predicts the future course of action of the phenomena. We can make future policies on the basis of estimates made with the help of Statistics. We can predict the demand for goods in 2005 if we know the population in 2004 on the basis of growth rate of population in past. Similarly a businessman can exploit the market situation in a successful manner if he knows about the trends in the market. The statistics help in shaping future policies.

6. Policy Making:

With help of statistics we can frame favourable policies. How much food is required to be imported in 2007? It depends on the food-production in 2007 and the demand for food in 2007. Without knowing these factors we cannot estimate the amount of imports. On the basis of forecast the government forms the policies about food grains, housing etc. But if the forecasting is not correct, then the whole set up will be affected.

7. It Enlarges Knowledge:

Whipple rightly remarks that “Statistics enables one to enlarge his horizon”. So when a person goes through various procedures of statistics, it widens his knowledge pattern. It also widens his thinking and reasoning power. It also helps him to reach to a rational conclusion.

8. To Measure Uncertainty:

Future is uncertain, but statistics help the various authorities in all the phenomenon of the world to make correct estimation by taking and analyzing the various data of the part. So the uncertainty could be decreased. As we have to make a forecast we have also to create trend behaviors of the past, for which we use techniques like regression, interpolation and time series analysis.

9.To simplify mass of figures:

Statistics helps in condensing mass of data into a few significant figures. Statistical methods present meaningful overall information from the mass of data. It is impossible to remember the individual income of the entire population. However remembering the figures of Percapita income is very easy.

Applications of Statistics:

https://prinsli.com/scope-and-importance-of-statistics/

Statistics has such a broad and ever-expanding scope that defining it is not only difficult but also unwise. Statistics are now used in practically every aspect of our life. It has become a recognized subject in its own right, as well as a tool of all disciplines essential to study, research and intelligent judgment. Statistical tools are used in almost every sector, including education, trade, industry or commerce, economics, biology, botany, medicine, physics, chemistry, astronomy, sociology, technology, and psychology.

Statistics has so many applications that it is often said, “Statistics is what statisticians do.” Governments, businesses, and people collect statistical data required in order to perform their activities efficiently and effectively. The following are some areas where statistics is used:

1. Statistics and State: (Importance of Statistics in State):

Since ancient times, ruling kings and chiefs have depended largely on statistics to formulate appropriate military and fiscal policies. Most of the statistics they collected, such as those on the population, military strength, crimes, taxes, and so on, were a by-product of administrative activity. The state's functions have greatly expanded in recent years. The concept of a state has evolved beyond just preserving law and order to one of providing social services.

2. Statistics in Business and Management: (Importance of Statistics in Business and Management):

Statistical methods are almost mainly used in the 20th century to solve commercial difficulties. Statistics are used in almost every aspect of business and industry, including manufacturing, financial analysis, distribution analysis, market research, personnel planning, research and development, and accounting, to name a few.

3. Statistics and Economics: (Importance of Statistics in Economics):

The study of statistics is an important element of economics. In 1890, the famous economist, Prof. Alfred Marshall, stated, "Statistics are the straw out of which I, like every other economist, have to create bricks." Dr Marshall's viewpoint highlights the importance of statistics in economics. Statistical data and statistical analysis techniques have been extremely helpful when dealing with economic issues. Prices, wages, time series analysis, and demand analysis are just a few examples.

4. Statistics and Physical Sciences: (Importance of Statistics in Physical Sciences):

Statistical methods were first developed and applied in the physical sciences, mainly physics, geology, and astronomy, but until recently, these sciences were not as involved in the development of 20th-century statistics as the biological and social sciences.

5. Statistics and Natural Sciences: (Importance of Statistics in Natural Sciences):

In the study of all natural sciences, like biology, medicine, meteorology, zoology, botany, etc., statistical methods have proven to be highly helpful. For example, the doctor had to rely heavily on factual data such as body temperature, pulse rate, BP, and so on when diagnosing the correct disease.

6. Statistics and Research: (Importance of Statistics in Research):

In research, statistics is a must-have. Most of the progress in knowledge has been due to experiments conducted with the help of statistical methods.

7. Statistics and Planning: (Importance of Statistics in Planning):

In the modern era, dubbed "the age of planning," statistics are essential to planning. Almost everywhere in the world, governments are restoring economic development planning.

8. Statistics and Industry: (Importance of Statistics in Industry):

Statistics are widely employed in industry to control inequality. In production engineering, statistical techniques such as inspection plans, control charts, etc., are used to determine if the product conforms to the standards or not. Other businesses use statistics to help them develop better products and services. Some companies use data from embedded sensors in their goods to provide services such as regular maintenance.

9. Statistics and Mathematics: (Importance of Statistics in Mathematics):

Statistics and mathematics are inextricably linked, and recent advances in statistical methods are the result of a wide range of mathematical applications.

Because statistics is a branch of applied mathematics, it is obvious that it plays an important role in mathematics. Statistics, on the other hand, is more than just a distinct branch of mathematics.

A large portion of math is focused on probability and theory. Statistical approaches help to improve the accuracy of such mathematical theories. Using averages, dispersion, and estimations, you can arrive at conclusions that are closer to the actual answer than simply guessing. Some examples of mathematical statistics are as follows:

- Calculate the average, median, and mode.
- In a group of 20 people, there is a 75% chance that at least two people will share the same birthday, etc.

10. Statistics and Medical Science: (Importance of Statistics in Medical Science):

In medical science, statistical tools for collecting, presenting, and analyzing observed data relating to disease reasons and incidence, as well as the effects of various medications and medicines, are extremely important.

11. Statistics and Psychology & Education: (Importance of Statistics in Psychology & Education):

Statistics has a wide range of applications in education and physiology, including identifying or attempting to determine the reliability and validity of a test, factor analysis, and so on.

12. Statistics and Other Uses: (Importance of Statistics in Other Fields):

Insurance companies, auditors, bankers, brokers, social workers, labor unions, trade unions, chambers of commerce, etc. can all benefit from statistics. Politicians and their supporters can benefit greatly from statistics. They want to see what their chances are and what efforts are required to win the election.

Limitations of Statistics:

1. Qualitative Aspect Ignored:
The statistical methods don’t study the nature of phenomenon which cannot be expressed in quantitative terms.

Such phenomena cannot be a part of the study of statistics. These include health, riches, intelligence etc. It needs conversion of qualitative data into quantitative data.

So experiments are being undertaken to measure the reactions of a man through data. Now a days statistics is used in all the aspects of the life as well as universal activities.

2. It does not deal with individual items:
It is clear from the definition given by Prof. Horace Sacrist, “By statistics we mean aggregates of facts…. and placed in relation to each other”, that statistics deals with only aggregates of facts or items and it does not recognize any individual item. Thus, individual terms as death of 6 persons

in a accident, 85% results of a class of a school in a particular year, will not amount to statistics as they are not placed in a group of similar items. It does not deal with the individual items, however, important they may be.

3. It does not depict entire story of phenomenon:

When even phenomena happen, that is due to many causes, but all these causes cannot be expressed in terms of data. So we cannot reach at the correct conclusions. Development of a group depends upon many social factors like, parents' economic condition, education, culture, region, administration by government etc. But all these factors cannot be placed in data. So we analyse only that data we find quantitatively and not qualitatively. So results or conclusion are not 100% correct because many aspects are ignored.

4. It is liable to be misused:

As W.I. King points out, "One of the short-comings of statistics is that do not bear on their face the label of their quality." So we can say that we can check the data and procedures of its approaching to conclusions. But these data may have been collected by inexperienced persons or they may have been dishonest or biased. As it is a delicate science and can be easily misused by an unscrupulous person. So data must be used with a caution. Otherwise results may prove to be disastrous.

5. Laws are not exact:

As far as two fundamental laws are concerned with statistics:

(i) Law of inertia of large numbers and

(ii) Law of statistical regularity, are not as good as their science laws.

They are based on probability. So these results will not always be as good as of scientific laws. On the basis of probability or interpolation, we can only estimate the production of paddy in 2008 but cannot make a claim that it would be exactly 100 %. Here only approximations are made.

6. Results are true only on average:

As discussed above, here the results are interpolated for which time series or regression or probability can be used. These are not absolutely true. If average of two sections of students in statistics is same, it does not mean that all the 50 students is section A has got same marks as in B. There may be much variation between the two. So we get average results.

“Statistics largely deals with averages and these averages may be made up of individual items radically different from each other.” —W.L King

7. To Many methods to study problems:
In this subject we use so many methods to find a single result. Variation can be found by quartile deviation, mean deviation or standard deviations and results vary in each case.

“It must not be assumed that the statistics is the only method to use in research, neither should this method of considered the best attack for the problem.” —Croxten and Cowden

8. Statistical results are not always beyond doubt:
“Statistics deals only with measurable aspects of things and therefore, can seldom give the complete solution to problem. They provide a basis for judgement but not the whole judgment.” —Prof. L.R. Connor

Although we use many laws and formulae in statistics but still the results achieved are not final and conclusive. As they are unable to give complete solution to a problem, the result must be taken and used with much wisdom.

CLASSIFICATION OF DATA:

Meaning of Classification of Data

- It is the process of arranging data into homogeneous (similar) groups according to their common characteristics.
- Raw data cannot be easily understood, and it is not fit for further analysis and interpretation. Arrangement of data helps users in comparison and analysis.
- For example, the population of a town can be grouped according to sex, age, marital status, etc.

Classification of data

The method of arranging data into homogeneous classes according to the common features present in the data is known as classification.

A planned data analysis system makes the fundamental data easy to find and recover. This can be of particular interest for legal discovery, risk management, and compliance. Written methods and sets of guidelines for data classification should determine what levels and measures the company will use to organise data and define the roles of employees within the business regarding input stewardship.

Once a data -classification scheme has been designed, the security standards that stipulate proper approaching practices for each division and the storage criteria that determines the data's lifecycle demands should be discussed.

Objectives of Data Classification

The primary objectives of data classification are:

- To consolidate the volume of data in such a way that similarities and differences can be quickly understood. Figures can consequently be ordered in sections with common traits.
- To aid comparison.
- To point out the important characteristics of the data at a flash.
- To give importance to the prominent data collected while separating the optional elements.
- To allow a statistical method of the materials gathered.

TYPES OF CLASSIFICATION:

There are four types of classification. They are

1. Geographical classification(Area wise e.g.: Cities, districts)
2. Chronological classification (Time)
3. Qualitative classification (Attributes)
4. Quantitative classification (Magnitudes)

(i) Geographical classification

When data are classified on the basis of location or areas, it is called geographical classification

Example: Classification of production of food grains in different states in India.

State wise estimates of Production of Food Grains:

Name of the State	Production of food grains(Thousand Tonnes)
Andhra Pradesh	1093.7
Bihar	12899.0
Haryana	11334.7
Punjab	21148.9
Uttar Pradesh	41828.6
All India	1,92,433.6

- Geographical classifications are listed in alphabetical order for easy reference.
- Items may be listed by size to emphasize the important areas as in ranking the States by Population.

(ii) Chronological classification

Chronological classification means classification on the basis of time, like months, years etc.

Example: Profits of a company from 2001 to 2005.

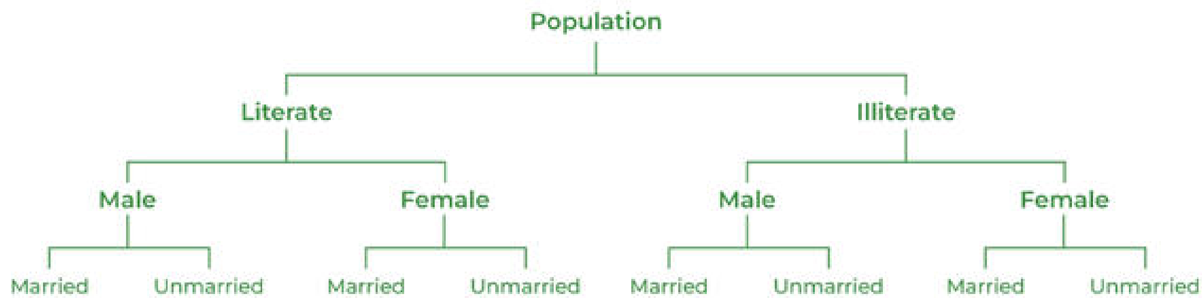
Population of India from 1951 to 2001

Year	Population(in Crores)
1951	36.11
1961	43.92
1971	54.82
1981	68.33
1991	84.63
2001	102.37

Time series are usually listed in chronological order, normally starting with the earliest period. When the major emphasis falls on the most recent events, a reverse time order may be used.

(iii) Qualitative classification

In Qualitative classification, data are classified on the basis of some attributes or quality such as sex, colour of hair, literacy and religion. In this type of classification, the attribute under study cannot be measured. It can only be found out whether it is present or absent in the units of study.



Classification of population on the basis of sex, i.e.into males and females or literacy i.e..into literate and illiterate. The type of classification where only two classes are formed is also called **two fold or Dichotomous Classifications.**

Instead of forming only two classes we further divide the data on the basis of some attributes so as to form several classes, the classification is known as **Manifold Classification.**

(iv) Quantitative classification

Quantitative classification refers to the classification of data according to some characteristics, which can be measured such as height, weight, income, profits etc.

Example: The students of a school may be classified according to the weight as follows.

Weight (in kgs)	**No of Students**
40-50	50
50-60	200
60-70	300
70-80	100
80-90	30
90-100	20
Total	700

There are **two types of quantitative classification of data.** They are

1. Discrete frequency distribution

2. Continuous frequency distribution

In this type of classification there are **two elements (i) variable (ii) frequency**

Variable

Variable refers to the characteristic that varies in magnitude or quantity. E.g. weight of the students. A variable may be discrete or continuous.

Discrete variable

- A discrete variable can take only certain specific values that are whole numbers (integers). E.g. Number of children in a family or Number of class rooms in a school.
- Varies with finite jumps.
- Manifests every conceivable fractional value.

No. of Children	**No. of Families**
0	10
1	40
2	80
3	100
4	250
5	150
6	50
Total	680

Continuous variable/ Continuous random Variable:

- A Continuous variable can take any numerical value within a specific interval.
- It is capable of manifesting every conceivable fractional value within the range of possibilities.

  Example: the average weight of a particular class student is between 60 and 80 kgs.

  Weight passes through all the values between these limits.

Weight(lbs)	**No. of persons**
100-110	10
110-120	15
120-130	40

130-140	45
140-150	20
150-160	04
Total	134

Frequency

Frequency refers to the number of times each variable gets repeated.

For example there are 50 students having weight of 60 Kgs. Here 50 students is the frequency.

Frequency distribution

Frequency distribution refers to data classified on the basis of some variable that can be measured such as prices, weight, height, wages, no. of units produced and consumed etc.

The following are the two examples of discrete and continuous frequency distribution

The following technical terms are important when a continuous frequency distribution is formed

Class limits: Class limits are the lowest and highest values that can be included in a class. For example take the class 40-50. The lowest value of the class is 40 and the highest value is 50. In this class there can be no value lesser than 40 or more than 50. 40 is the lower class limit and 50 is the upper class limit.

Class interval: The difference between the upper and lower limit of a class is known as class interval of that class. Example in the class 40-50 the class interval is 10 (i.e. 50 minus 40).

Class frequency: The number of observations corresponding to a particular class is known as the frequency of that class

Example:

Income (Rs) **No. of persons**

1000 - 2000 50

In the above example, 50 is the class frequency. This means that 50 persons earn an income between Rs.1, 000 and Rs.2, 000.

(iv) Class mid-point: it is the value lying half way between the lower and upper class limits of a class interval.

the

band.

5. Ratio graph

It is also called semilogarithmic graph.

The absolute changes in the values of a variable from one period to another can be shown on natural (or) arithmetic scale.

when relative rates of change are to be studied logarithmic (or) ratio scale is used.

In ratio scale equal vertical distances indicate equal relative rates of change (or) equal percentage.

The Natural scale is based on arithmetic progression where as the ratio scale is based on geometric progression.

The ratio chart (or) ratio scale enable us to compare the rate of change of categories of different statistical units on the same chart. Many curves can be plotted on the same graph and their trends studied.

In case of ratio scale, the Y-axis start from one and not from zero whereas in the case of the scale on semi-logarithmic paper starting at 1 and not zero is that the logarithm of 1 is 0. Hence 1 is placed at zero distance from the origin.

There is no logarithm for zero, nor for -ve numbers, hence such values cannot be plotted.

On the logarithmic scale chart, constant rate of increase or decrease can be easily noticed and this property is of great use.

In case of variables having wide range of values the ratio scale graph is far more suitable than the other.

In ratio scale the meaning of the data is derived from the direction of lines whereas in case of natural scale the meaning is derived from position of lines.

Methods of constructing a semi-logarithmic graph

A semi-logarithmic graph can be constructed in any of the following ways.

1. By plotting the logarithmic of the given values on a natural scale.

2. By plotting the given values on a semi-logarithmic paper.

when first method is adopted the logarithm's of the various values of variable are obtained by consulting the logarithmic tables. These logarithms are then plotted on Y-axis of the natural scale and the various points are joined by straight lines.

when second method is adopted, no need to find log values to the variables, rather the actual values are plotted on semi-logarithmic paper. This method is simple and convenient as compared to the first method.

when a graph is prepared by following any of the above two methods, it is known as semi-logarithmic graph, here vertical scale is ruled on the ratio principle but the horizontal scale remains on the arithmetic principle.

It is also possible to have a line graph that has both the scales logarithmic, such a graph is known as a double logarithmic graph and the use of this graph is limited.

Interpretation of logarithmic curves :-

The logarithmic curves must be interpreted with caution, otherwise there is a possibility of jumping to wrong conclusions, The following are some of the important points.

1. If a curve is raising upwards, it would indicate an increasing rate of change.

2. If the curve is falling downwards, it represents a decreasing rate of change.

3. If a curve is a straight line, the rate of change is constant (or) uniform.

4. If a curve is rising but is nearly straight, it represents a decline at a nearly uniform rate.

5. If a curve is steeper in one portion than in another portion, the rate of change in the former is more rapid than that in the latter.

6. If two curves on the same ratio chart are found running parallel, they represent equal percentage of change.

7. If one curve is steeper than another on the same ratio chart, the rate of change in the former is more rapid than that in the latter.

Uses of ratio chart :-

Ratio graphs are useful for 4 types of comparision.

1. A constant percent rate of growth is represented by a straight line such as the sales increasing 10 percent a year appear on the ratio chart as straight line. If the series curves are away from the straight line, it denotes a corresponding change in the rate of growth.

2. If historic factors of growth may be expected to persist, the analyst can project past trends, in order to forecast future volumes.

3. The relative growth of fluctuation of two curves may be compared more accurately in ratio charts than in arithmetic charts since parallel lines indicates the same percent rates of change anywhere on the chart and steeper slopes indicate higher rates.

4. percentages (or) ratios may be read directly from the vertical scale and applied towards further graphic analysis.

5. Ratio scale is extremely useful in comparing series which differ widely in magnitude.

Limitations:-

1. They are difficult to understand.
2. Zero (or) Negative values cannot be shown.
3. The study of an aggregate into various component parts is not possible by using ratio scale.
4. Their interpretation needs highly specialised knowledge in the absence of which one may draw entirely wrong conclusions. This factor alone restricts the scope of mass popularly of such a useful
5. Ratio scale cannot measure absolute changes.

Types of frequency distribution Graphs:-

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

1. Histograms
2. frequency polygon
3. Smoothed frequency curve
4. ogives (or) cumulative frequency curves.

1. Histograms:-

Histogram is most widely used for graphical representation of a frequency distribution. while constructing histogram the variable is always taken on x-axis and frequencies depending on it (or) on y-axis. The area of the histogram represents the total frequency as distributed through out the classes. Histogram is a two dimensional i.e., length and width are important.

The technique of constructing histogram is

(i) for distributions have equal class intervals.

(ii) for distributions having unequal class intervals.

when class intervals are equal, take frequency on y-axis, the variable on x-axis and construct adjacent rectangles. In such case height of the rectangles will be proportional to the frequencies.

when class intervals are unequal, a correction for unequal class intervals must be made. The correction consists of find a each class the frequency density (or) the relative frequency density. The frequency density is the frequency for that class divided by the width of the class.

A Histogram (or) frequency density polygon constructed from these density values would have the same general appearance as the corresponding graphical display developed from equal class interval.

For adjustment take the class which has lowest class - interval and adjust the frequencies of other classes in the following way. If one class interval is twice as wide as the one having lowest class interval, divide the height of the rectangle by two. If it is three time more, divide the height of it rectangle by three etc... i.e., the heights will be porportional to the ratios of the frequencies of the width of the class.

2. Frequency polygon:-

A frequency polygon is a graph of frequency distribution. It has more than four sides. It is particularly effective in comparing 2 or more frequency distributions. There are 2 ways to construct a frequency polygon.

1. Draw a histogram of the given data and then join by straight lines the midpoints of the upper horizontal side of each rectangle with the adjacent ones. The figure so obtained is called frequency polygon. practice to close the polygon at both ends of the distribution by extending them to base line of making the area under polygon equal to the area under the corresponding histogram.

2. Another method is to take the class midpoints of the various class intervals and then plot the frequency corresponding to each point and join these points by straight lines. The figure so obtained is same as by the model 1 but the only difference is there here we have not to construct histogram.

Advantages:-

1. In a frequency polygon the values of made can be easily obtained if from the apex of a Polygon a far is drawn on x-axis.

2. It facilitates comparision of 2 (or) more frequency distribution on the same graph.
3. Frequency polygon is simpler than its histogram.
4. It sketches an outline of the data pattern more clearly.
5. Polygon becomes increasingly smooth and curve like as we increase the number of classes and no. of observation.

Same as in histogram difficulties are faced in the construction of a frequency polygon i.e., they cannot be used for distributions having open end classes. To avoid it make an adjustment same as in histogram.

3. 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 same as of polygon. The objective is to eliminate the accidental variations present in the data. For drawing a smoothed frequency curve it is necessary to first draw the polygon and then smooth it out, while constructing polygon by plotting the frequencies at the mid point of class intervals save some time but the smoothing of the polygon cannot be done properly without a histogram.

i.e, first draw histogram then polygon and lastly smoothen it. The curve should begin and end at the base line and as a general rule it may be extended to the midpoints of the class intervals just outside the histogram.

The area under the curve should represents the total number of frequencies in the entire distribution.

The following points are important while smoothing a frequency curve.

1. only frequency distribution based on samples should be smoothed.
2. only continuous seried should be smoothed.
3. The total area under the curve should be equal to the area ~~under~~ under the original histogram (or) polygon.

4. Cumulative frequency curves (or) ogives :-

It is a graphical representation of cumulative frequency distribution. An ogive is obtained by plotting the cumulative frequency on y-axis against the class boundaries on x-axis.

Types of ogive curves : There are 2 types

1. Less than ogive
2. More than ogive

1. Less than ogive :-

It is constructed on the basis of cumulative frequencies which are in ascending order. In less than ogive, cumulative frequencies are plotted against upper limits of respective class. It is an increasing curve which has an upward slope from left to right.

2. More than ogive:-

It is constructed on the basis of cumulative frequencies which are in descending order. In more than ogive, cumulative frequencies are plotted against lower limits of respectives class. It is a decreasing curve which has downwards slope from left to right.

Uses of ogives:-

1. It helps to determine graphically the number of proportion observation above (or) below the given value of variable.
2. It also helps to complete partition values such as median and quartiles graphically.
3. It facilitates the comparision of two frequency distributions.

Differences between Histogram and Histrogram

Histogram shows vertical adjacent rectangles with class intervals on x-axis and corresponding frequencies on y-axis. Histogram shows changes in variable over a period of time with time periods on x-axis and values of variable on y-axis.

For example: Scale (or) population ~~of variable~~ over a period of 10 years from 2001 to 2010

Basics of distinction	Histogram	Historigram
1. curve/diagram	It is an area diagram.	It is a frequency curve
2. Type of Graph	It is a graph of frequency distribution.	It is a graph of time series.
3. x-axis	It shows class intervals on x-axis	It shows time-period on x-axis.
4 y-axis	It shows frequencies of class intervals on y-axis.	It shows the values of a variable on y-axis.

Unit-3

Measures of Central Tendency and dispersion

Introduction:

An average is a single value which is used to represent all of the values in the series. The average lies in between two Extremes. That is largest and smallest items. It is some times called as measure of "Central Tendency".

Objective of Central Tendency:

-> To determine one single value that may be used to discribe the characteristics of Entire Series.

-> To facilitates comparission

-> To facilitates statistical inference

-> To help in decission making process

Characteristics of an ideal average :-

-> It should be Easy to understand.

-> It should be simple and comparable.

-> It Should be based on all the items of the serie's

-> It should not be affected by Extreme items.

-> It should be rigidly defined by a mathematical formula.

-> It should be capable of further and algebric treatment

-> It should have sampling stability.

Definition of an average:

Average is an attempt to find one single figure to describe whole of the figures.

Unit-3

Measures of Central Tendency and dispersion

Introduction:-

An average is a single value which is used to represent all of the values in the series. The average lies in between two Extremes. That is largest and smallest items. It is some times called as measure of "Central Tendency".

Objective of Central Tendency:

-> To Determine one single value that may be used to discribe the charectaristics of Entire Series.

-> To facilitates comparission

-> To facilitates statistical inference

-> To help in decission making process

Charectaristics of an ideal average:-

-> It should be Easy to understand.

-> It should be simple and comparable.

-> It should be based on all the items of the serie's

-> It should not be affected by Extreme items.

-> It should be rigidly defined by a mateametical formula.

-> It should be capable of further all algebric treatment

-> It should have sampling stability.

Definition of an average:

Average is an attempt to find one single figure to describe whole of the figures.

Types of Averages:

1. Arteamatic mean

   i) simple artuematic mean
      - indivisual observation
      - descorete series

   ii) weighted artuamatic mean

2. Median

3. Mode

4. Geomatoric mean

5. Harmonic mean.

1. Arteamatic mean:

The most popolar and vitly used measure of representing the entire data by one value is called as an average and statisticians call it as "artuematic mean".

It is uptained by adding togetuer all tue items and by dividing tuis totals by tue noumber of items

simple artue matic mean:-

indivisual Absorvation:

$$\overline{x} = \frac{x_1 + x_2 + x_3 - - - - - + x_n}{N}$$

$$\overline{x} = \frac{\Sigma x}{N}$$

Where $\overline{x}$ = Arteamatic mean

Σx = Som of all values of x variable ($x_1 + x_2 + x_3 - - - x_n$)

N = Noumber of observations

step:1

Add together all the values of variable x and explain the total that is Σx

step-2. Divide the total by the number of observation by the N

The following table gives the monthly income of 10 Employees in an office.

incomes: 14780; 15760; 26690; 27750; 24840; 24920; 19100; 17810; 27050; 26950

calculat A.M

Employee	income
1	14780
2	15760
3	26690
4	27750
5	24840
6	24920
7	19100
8	17810
9	27050
10	26950
N = 10	Σx = 2,25,650

$$\bar{x} = \frac{\Sigma x}{N}$$

$$\bar{x} = \frac{2,25,650}{10}$$

$$\therefore \bar{x} = 2.25.65\cancel{0}-$$

Hence the average of 10 Employee is ∴ 2.25, 650-----

Calculate mean from the following data:

Income p.A (in lakhs): 10. 20. 30. 40. 50. 60. 70. 80

Employee	income data
1	10
2	20
3	30
4	40
5	50
6	60
7	70
8	80
N = 8	$\Sigma x = 360$

$$\bar{x} = \frac{\Sigma x}{N}$$

$$\bar{x} = \frac{360}{8}$$

$$\bar{x} = 45$$

Hence the average income is

∴ 45,00,000/-

Shortcut cut method:-

$$\bar{x} = A + \frac{\Sigma d}{N}$$

where A = Assumed mean

d = deviations of assumed means

Deviations of items assumed mean

$d = (x - A)$

Steps;-

Take an assumed mean

Step 2 Take the deviations of items from the assumed mean

Step 3 Obtain the sum of these deviations

Step 4 apply the formula.

calculate the Arithmatic mean by shortcut method

Employee	Income	$d = x - A$
1	14780	14780 - 24920 = -10.140
2	15760	15760 - 24920 = -9160
3	26.690	26690 - 24920 = 1770
4	27.750	27.750 - 24920 = 2830
5	24840	24840 - 24920 = -80
6	$\boxed{24920}$ M	24920 - 24920 = 0
7	19.100	19100 - 24920 = -5820
8	17.810	17810 - 24920 = -7110
9	27.050	27050 - 24920 = 2130
10	26,950	26950 - 24920 = 2030
N = 10	= 2,25,650	Σd = 23550

8760

32310

-23550

$\bar{x} = A + \frac{\Sigma d}{N}$

$= 24920 + \frac{23550}{10}$

= 24920 + 2355

= 22.565 where Arithmaticmean

From the following data calculate A.M by direct and and short cut method.

Roll	1	2	3	4	5	6
mark	5	15	25	35	45	55

Roll	Marks
1	5
2	15
3	25
4	35
5	45
6	55
N = 6	Σx = 180

$$\bar{x} = \frac{\Sigma x}{N}$$

$$\bar{x} = \frac{180}{6}$$

$$\bar{x} = 30/-$$

∴ Where the Arthamatic Mea

Roll	Marks	d = x-A
1	5	-20
2	15	-10
3	25 = A	0
4	35	10
5	45	20
6	55	30
N = 6		Σd = 30

$$\bar{x} = A + \frac{\Sigma d}{N}$$

$$\bar{x} = 25 + \frac{30}{6}$$

$$\bar{x} = 25 + 5$$

$$\bar{x} = 30$$

$\boxed{\therefore AM = 30}$

Calculation of A.m in discrete series.

Direct method:

$$\bar{x} = \frac{\Sigma fx}{N}$$

f = frequency

x = Variables

N = Total number of observations

Step 1 :-

Multiply the frequency of each row with a variable and optain the total Σfx

Step 2 divide the total optained by step 1 by the number of observations.

Calculate A.M for the following data

Marks :	20	30	40	50	60	70
No. of students	8	12	20	10	6	4

Marks	No. of stu	Σfx
20	8	$20 \times 8 = 160$
30	12	$30 \times 12 = 360$
40	20	800
50	10	500
60	6	360
70	4	280
	$N = 60$	$\Sigma fx = 2460$

$$\bar{x} = \frac{\Sigma fx}{N}$$

$$\bar{x} = \frac{2460}{60}$$

$$\boxed{\bar{x} = 41}$$

Short cut method:

$$\bar{x} = A + \frac{\Sigma fd}{N}$$

Where A = Assumed mean

f = frequency

d = diviation of item from the assumed mean

N = Total number of observations.

Step:1 Take an assumed mean

Step 2 Take the deviations of the variable x from the assumed

Step 3 Multiply these deviations with the respective frequency and take the total Σfd

calculate AM by Short cut method

Marks (x)	No. of student (f)	$d = x - A$	Σfd
20	8	-20	8×20 = -160
30	12	-10	-120
40 (A)	20	0	0
50	10	10	100
60	6	20	120
70	4	30	120
	N: 60	d = 30	Σfd = 60

$\bar{x} = A + \frac{\Sigma fd}{N}$

$\bar{x} = 40 + \frac{60}{60}$

$\bar{x} = 40 + 1$

$\boxed{\bar{x} = 41}$

Calculate AM for discrete series in direct and short cut method.

Marks	5	15	25	35	45	55
No. of students	10	20	30	50	40	30

direct method:

$\bar{x} = \frac{\Sigma fx}{N}$

$\bar{x} = \frac{6300}{180}$

$\bar{x} = 35$

Shortcut method:

$\bar{x} = A + \frac{\Sigma fd}{N}$

$\bar{x} = 25 + \frac{1800}{180}$ $\frac{1800}{180}$

$\bar{x} = 25 + 10$

$\bar{x} = 35$

(x) Marks	(f) No.of Students	fx	d = x-A	fd
5	10	50	-20	-200
15	20	300	-10	-200
25 (A)	30	750	0	0
35	50	1750	10	500
45	40	1800	20	800
55	30	1650	30	900
	N = 180	$\Sigma fx = 6300$	d = 30	Σfd = ~~1050~~ 1800

Calculation of Arithmetic Mean in continuous method

$$\overline{x} = \frac{\Sigma fm}{N}$$

∴ where f = frequency

m = mid point various classes

N = Total number of observations

Step 1 :- obtain midpoint of each class (m)

Step 2 : multiply these mid points by the respective frequency of each class and obtain the total Σfm

Step 3 : Divide the total obtained in step 2 by the sum of frequency

$$\text{midpoint (m)} = \frac{\text{upper limit} + \text{lower limit}}{2}$$

from the following data compute AM Method by direct method.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No.of students	5	10	25	30	20	10

Marks	No.of students	Midpoint	fm
0-10	5	5	25
10-20	10	15	150
20-30	25	25	625
30-40	30	35	1050
40-50	20	45	900
50-60	10	55	550
	N = 100		$\Sigma fm = 3300$

$\bar{x} = \frac{\Sigma fm}{N}$

$\bar{x} = \frac{3300}{100}$

$\bar{x} = 33$

Shortcut method:

$$\bar{x} = A + \frac{\Sigma fd}{N}$$

A: Assumed mean

f: frequency

d: diviation of mid points from assumed mean

$d = M - A$

N: Total number of observation

Step 1: Take an assumed mean

Step 2: Take the diviation of mid points from assumed mean

Step 3: multiply the respective frequency of each class by the diviations and obtain the total Σfd

Apply the formula

Marks	No. of std (f)	Midpoint $\frac{U+L}{2}$	d = M - A	fd
0-10	5	5	-30	-150
10-20	10	15	-20	-200
20-30	25	25	-10	-250
30-40	30	35 (A)	0	0
40-50	20	45	10	200
50-60	10	55	20	200
	N = 100			fd = -200

600

400

-200

$$\bar{x} = A + \frac{\Sigma fd}{N}$$

$$35 + \frac{-200}{100} = 35 + (-2) = 35 - 2 = 33$$

$$\boxed{\therefore \bar{x} = 33}$$

In order to specify calculations divide the deviations by class interval's that is calculate $\frac{M-A}{i}$ and then multiply by i. the formula is as follows

$$\bar{x} = A + \frac{\Sigma fd}{N} \times i$$

A = Assumed mean

f = frequency

d = deviation of midpoints from assumed mean

$$\frac{m - A}{i}$$

N = Total no. of observations

i = class interval

- 0.2

- 2

Marks	No. of. stu	Midpoint $\frac{U+l}{2}$	$d = \frac{m-A}{i}$	Σfd
0-10	5	5	$\frac{5-35}{10} = \frac{-30}{10} = -3$	-15
10-20	10	15	$\frac{15-35}{10} = -2$	-10
20-30	25	25	-1	-25
30-40	30	35 (A)	0	0
40-50	20	45	1	20
50-60	10	55	2	20
	N = 100			$\Sigma fd = -20$

$\bar{x} = A + \frac{\Sigma fd}{N} \times i$

$= 35 + \frac{-20}{100} \times 10$

$\bar{x} = 35 + (-0.2) \times 10$

$\bar{x} = 35 - 2$

$\boxed{\bar{x} = 33}$

Calculate AM by direct and shoortcut metuod

marks less than	10	20	30	40	50	60
No of stu	10	30	60	110	150	180

Marks less than	No. of. stud	Midpoint = $\frac{U+l}{2}$ d	fd = mA
0-10	10	5	50
10-20	30	15	450
20-30	60	25	1500
30-40	110	35	3850
40-50	150	45	6750
50-60	180	55	9900
	N = 540		fd = 22500

$\bar{x} = \frac{\Sigma fM}{N}$

$\bar{x} = \frac{22500}{540}$

$\boxed{\bar{x} = 41.66}$

Graph:

i = class intervals

Marks	No. of Stu	Mid point	$d = \frac{m-A}{i}$	fd.
0-10	10	5	$\frac{5-25}{10} = -2$	-20
10-20	30	15	$\frac{15-25}{10} = -1$	-30
20-30	60	25 (A)	0	0
30-40	110	35	1	110
40-50	150	45	2	300
50-60	180	55	3	~~360~~ 540
	N = 540			fd = ~~720~~ 900

$$\bar{x} = A + \frac{\Sigma fd}{N} \times i$$

$$\bar{x} = 25 + \frac{900}{540} \times 10$$

$$\bar{x} = 25 + 1.66 \times 10$$

$$\bar{x} = 25 + 16.6$$

$$\boxed{\bar{x} = 41.66}$$

~~$\bar{x} = 25 + 1.33 \times 10$~~

~~$\bar{x} = 25 + 13.33$~~

calculate AM for the following continuous series

Marks	0-10	10-30	30-60	60-100
No of stu	5	12	25	08

$\bar{x} = A + \frac{\Sigma fd}{N} \times C$ C: common factor

Marks	(f) No. of stu	Midpoint $\frac{U+L}{2}$	$d = \frac{m-A}{c}$	fd
0-10	5	5 (c)	$\frac{5-45}{5} = -8$	-40
10-30	12	20	$\frac{20-45}{5} = -5$	-60
30-60	25	45 (A)	$\frac{45-45}{5} = 0$	0
60-100	08	80	$\frac{80-45}{5} = 7$	56
	N = 50			$\Sigma fd = -44$

$\bar{x} = A + \frac{\Sigma fd}{N} \times C$

$\bar{x} = 45 + \frac{-44}{50} \times 5$

$\bar{x} = 45 + (-4.4)$

$\boxed{\bar{x} = 40.6}$

Correcting - Incorrect Items:

The mean marks of 100 student's was found to be 40 later on it was found that the score of 53 was misread as 83. find the correct mean corresponding to the correct score.

($\bar{x}$) Mean = 40

N = 100

$\bar{x} = \frac{\Sigma x}{N}$

$\bar{x} = \frac{\Sigma x}{N} = 40 = \frac{\Sigma x}{100}$

$= 40 \times 100$

$\Sigma x = 4000/-$

It is incorrect.

Correct Σx = Incorrect Σx - wrong Num + Correct Num.

$= 4000 - 83 + 53$

$\Sigma x = 4000 - 30$

$\Sigma x = 3970$

$\bar{x} = \frac{\Sigma x}{N}$

$= \frac{3970}{100} = 39.7$

mean of 100 observations is found to be 40. if at the time of computation Two items were wrongly taken as 30 and 27 instead of 3 and 72 find the correct mean.

$\bar{x} = 40$

$N = 100$

$\bar{x} = \frac{\Sigma x}{N}$

$= 40 \times \frac{\Sigma x}{100}$

$= 4000$

Where it is incorrect.

① Correct Σx = Incorrect - wrong num + correct num

$= 4000 - 30 - 27 + 3 + 72$

$= 4000 - 30 + 27 + 75$

$= 4000 - 57 + 75$

$= 4000 + 18$

$= 4018$

$\bar{x} = \frac{\Sigma x}{N}$

$= \frac{4018}{100}$

$= 40.18$

Merits : Advantages:

-> It is Easy to understand and easy to calculate.

2. It is based on all the observations.

3. It is rigidly displain define

4. It is the center of gravity that balances the values on a Either side of the meter.

5. It is relatively reliable

Demerits:

1. It can't be determined by inspiration

2. It can't be located practically

3. It can't use qualitative characteristics like honesty, unity

and intelligence.

5. It is affected very much by Extreme values

6. It can't be optained by a single observation.

Weighted arthamatic mean

$$\bar{x} = \frac{\Sigma x w}{\Sigma w}$$

if a candidate optains the following marks English-75, stats-60, maths-59, physics-55, Chemistry-63, find the weighted mean if weights 2, 1, 3, 3, and 1 respectively are alloted to the subjects.

Subject	Marks (x)	weights w	xw
English	75	2	150
stats	60	1	60
maths	59	3	177
Physics	55	3	165
Chemistry	63	1	63
		$\Sigma w = 10$	$\Sigma xw = 615$

$$\bar{x} = \frac{\Sigma x w}{\Sigma w}$$

$$\bar{x} = \frac{615}{10}$$

$$\boxed{\bar{x} = 61.5}$$

calculate simple artuemetic mean and weighted artuamatic mean for the following data.

Course	(x) pas spercetge	w No.of.stu	xw
MCA	71	3	213
MBA	83	4	332
Mcom	73	5	365
BA	74	2	148
Bcom	65	3	195
BSC	66	3	198
N = 6	432	20	$\Sigma xw = 1451$

$\bar{x} = \frac{\Sigma x}{N}$

$= \frac{432}{6}$

$= 72$

$\bar{x} = \frac{\Sigma xw}{\Sigma w}$

$= \frac{1451}{20}$

$= 72.5$

from the following data calculate simple Arithmatic and weighted arithmatic mean for the colleges A, B, C and find out the performance of best college

College course	A No.%	A No of Students	B Pass%	B No of Stud	C Pass%	C No of Stud
MA	71	3	82	2	81	2
Mcom	83	4	76	3	76	3.5
BA	73	5	73	6	74	4.5
B.com	74	2	76	7	58	2
Bsc	65	3	65	3	70	7
Msc	66	3	60	7	73	2

	A			B			C		
Cource	(x) No.of.%	(w) No.of.stu	ΣxW	(x) No.of.%	(w) No.of stu	ΣxW	(x) No.of per%	(w) No.of.stu	ΣxW
MA	71	3	213	82	2	164	81	2	162
M.com	83	4	332	76	3	228	76	3.5	266
BA	73	5	365	73	6	438	74	4.5	333
B.com	74	2	148	76	7	532	58	2	116
BSC	65	3	195	65	3	195	70	7	490
MSE	66	3	198	60	7	420	73	2	146
N = 6	x = 432	w = 20	Σxw = 1451	x = 432	w = 28	xw = 1977	x = 432	w = 21	xw = 1513

A

$\bar{x} = \frac{\Sigma x}{N}$ $\qquad$ $\bar{x} = \frac{\Sigma xw}{\Sigma w}$

$\bar{x} = \frac{432}{6}$ $\qquad$ $\bar{x} = \frac{1451}{20}$

$\bar{x} = 72$ $\qquad$ $\bar{x} = 72.5$

B

$\bar{x} = \frac{\Sigma x}{N}$ $\qquad$ $\bar{x} = \frac{\Sigma xw}{\Sigma w}$

$\bar{x} = \frac{432}{6}$ $<$ $\bar{x} = \frac{1977}{28}$

$\bar{x} = 72$ $\qquad$ $\bar{x} = 70.6$

C

$\bar{x} = \frac{\Sigma x}{N}$ $\qquad$ $\bar{x} = \frac{\Sigma xw}{\Sigma w}$

$\bar{x} = \frac{432}{6}$ $\qquad$ $\bar{x} = \frac{1513}{21}$

$\bar{x} = 72$ $\qquad$ $\bar{x} = 72.0$

Performence of Best College is A

The weighted A.M for college A is higher than college B & C $\therefore$ college A is considred as best college.

Median

Median is a value of the variable which divides data into two Equal parts. it reffers to the middle value of a distribution

calculation of Median Individual series:

step 1: Arrange the size of in accending order and decending order. Apply $\left(\frac{N+1}{2}\right)^{th}$ item calculate median.

From the following data of wages of 7 workers calculate the median wage.

wage in (₹) 1100, 1150, 1080, 1120, 1200, 1160, 1400

S.NO	Accending order
1	1080
2	1100
3	1120
4	1150
5	1160
6	1200
7	1400

Median = $\left(\frac{n+1}{2}\right)^{th}$ item

$M = \frac{7+1}{2} = 8/2$

$M = 4$

$\boxed{\therefore \text{Median} = 4}$

size of 4th item is 1150

$\boxed{\therefore \text{Median wage} = 1150}$

uptain the value of median from the following data.

391, 384, 591, 407, 672, 522, 777, 753, 2488, 1490

NO	Accending order	S.NO	Accending order
1	384	6	672
2	391	7	753
3	407	8	777
4	522	9	1490
5	591	10	2488

Median = $\left(\frac{n+1}{2}\right)^{th}$ item

$M = \frac{10+1}{2} = 5.5$ item

size of 5.5 = 5th + 6th item.

$= \frac{591+672}{2} = \frac{1263}{2}$

$\boxed{\therefore \text{Median} = 631.5}$

Calculation of Median discrete series

Step 1: Arrange the size of items in Assending order (or) decending order

2: Calculate cumulative frequency

3: Apply $\frac{n+1}{2}$ formula

4: find out the cumulative frequency which includes $\left(\frac{n+1}{2}\right)^{th}$ item calculate median.

Median: Size of items corresponding to the cumulative frequency which includes $\left(\frac{n+1}{2}\right)^{th}$ item.

Marks	45	55	25	35	5	15
No.of stu	40	30	30	50	10	20

Marks	(f) No.of students	C.f
5	10	10
15	20	30
25	30	60
35	30	90
$\boxed{45}$	40	(130)M
55	50	180.
	N = 180	

$M = \left(\frac{n+1}{2}\right)^{th}$ item

$= \frac{180+1}{2} = \frac{181}{2} = 90.5$

$\therefore$ 90.5 item = 130

Median, 130 of c.f lies in the marks of 45

from the following data find out the value of median.

Income	1000	1500	800	2000	2500	1800
No.of persons	24	26	16	20	06	30

Income	No of persons	C.f
800	6	6
1000	16	22
1500	20	42
1800	24	66
2000	26	92
2500	30	122
	$\Sigma f = 122$ $N =$	

Median, $\left(\frac{N+1}{2}\right)^{th}$ item

$= \frac{122+1}{2} = \frac{133}{2}$

$= 61.5$

M ∴ 61.5 item is = ~~1800~~ 66

Median 66 of c.f in the income = 1800

Calculation of median continuous series

Step 1: calculate cumulative frequency

2: Ascertain cumulative frequency which includes $\left(\frac{N}{2}\right)^{th}$ item. the corresponding and lower limit (L) of the class and (I) of the class interval of the corresponding class and cumulative frequency of the preceding ~~class~~ class

3: calculate median

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

$$L + \frac{\frac{N}{2} - cf}{f} \times i$$

L = lower class

cf = cumulative frequency

f = frequency of the class

i = class interval.

N = Σf = Total number of frequency.

-> class intervals only arranging in ascending order.

Calculate Median

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No.of Students	10	20	30	50	40	30

Marks	No.of. Students	C.f
0-10	10	10
10-20	20	30
20-30	30	60 cf
30-40	50 f	110
40-50	40	150
50-60	30	180
	N = 180	

$\left(\frac{n}{2}\right)^{th} = \frac{180}{2} = 90$

$M = L + \frac{\frac{N}{2} - cf}{f} \times i$

$M = 30 + \frac{90 - 60}{50} \times 10$

$M = 30 + \frac{30}{50} \times 10$

$M = 30 + 0.6 \times 10$

$M = 30 + 6$

$\boxed{\therefore \text{Median} = 36}$

Calculate Median for the following frequency distribution.

Marks	45-50	40-45	35-40	30-35	25-30	20-25	15-20	10-15	5-10
No.of stud	10	15	26	30	42	31	24	15	07

Marks	No.of.stu	C.f
5-10	7	7
10-15	15	22
15-20	24	46
20-25	31	77 Cf
25-30	42 f	119
30-35	30	149
35-40	26	175
40-45	15	190
45-50	10	200
	N = 200	

$\left(\frac{n}{2}\right)^{th} = \frac{200}{2} = 100$

$M = L + \frac{\frac{n}{2} - cf}{f} \times i$

$M = 25 + \frac{100 - 77}{42} \times 5$

$M = 25 + \frac{23}{42} \times 5$

$M = 25 + 0.5 \times 5$

$M = 25 + 2.5$

$M = 27.5$

Merits of Median

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. It is not at all affected by extreme values.
4. It can be calculated for distributions with open ended claus
5. Median is the only average to be used while dealing with qualitative data.
6. It can be determined graficaly.

Demerits of median

1. In case of Even number of observations median can not be determined Exactly.
2. It is not based on all the observations.
3. It is not capable of further mathamatic treatment.

Mode

It is the value in a series of observations which occurs with greatest frequency

Individual series

find mode for the following data.

3 , 5. 8. 5, 4 . 5. 9. 3 (unimodel)

5 is repeated more time

∴ mode = 5 (unimode)

find mode for the following data

2. 2. 4, 6; 8. 10, 12, 12, 14, 16.

∴ Mode = 2 and 12

(Bi mode)

Descreat series

find the mode from the following frequency distribution.

Size	1	2	3	4	5	6	7	8	9	10	11	12
frequency	3	8	15	23	35	40	32	28	20	45	14	06

steps for grouping table:

Step 1 : The frequency of coloumn 1 are the orginal frequency's

Step 2 : Coloumn is uptained by combining the frequencys two by two

Step 3 : in coloumn 3 leave the first frequency and combine the remaining frequency two by two.

Step 4 : in coloumn 4 combined the frequencys three by three.

Step 5 : in coloumn leave the first frequency and combine remaining frequency three by three

UNIT-IV

Correlation

Introduction:

Correlation is the relationship that exists b/w two (or) more variable i.e., if two variables are related one is changed the other will automatically be changed.

ex: Heights & weights, Demand & Supply, price & Demand Etc.,

Definition:

"Correlation analysis is a statistical technique used to measure the degree and direction of relationship between variables".

Significance of Correlation:

→ Economic theory and business studies show relationship between variables.

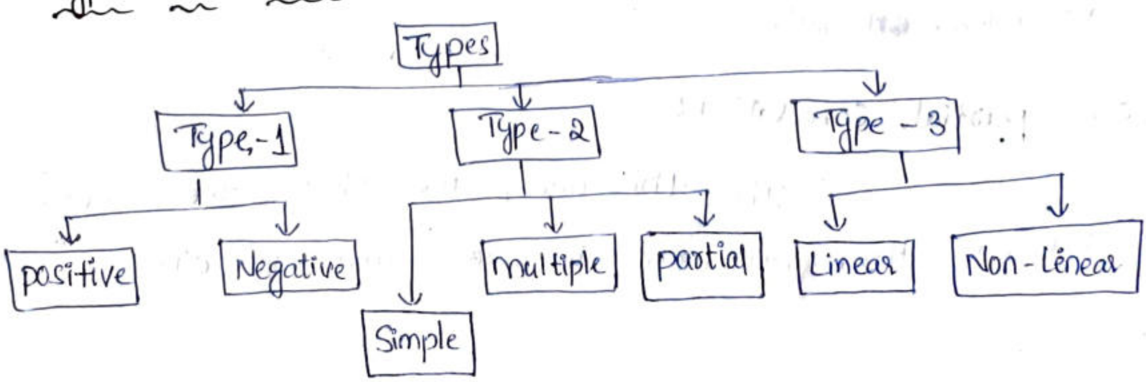
→ Correlation analysis helps in deriving exactly the degree and direction of such relationships.

→ Correlation analysis contributes to the understanding of Economic behaviour.

→ The measure of co-efficient of correlation is relative measure of change.

→ The Effect of correlation is to reduce range of uncertainity of our prediction.

Types of Correlation:



01. **positive correlation :**

If both the variables vary in same direction is called positive correlation i.e., if one variable increases, the other also increases (or) if one variable decreases the other also decrease.

Ex :-

x	10	20	30	40	50
y	1	2	3	4	5

02. **Negative Correlation :**

If both the variables vary in opposite direction, i.e., called Negative Correlation i.e., if one variable increases, the other variable decreases. (or) if one variable decreases, the other variable increases.

Ex :-

x	10	20	30	40	50
y	10	8	6	4	2

03. **Simple Correlation :**

When only two variables are studied i.e., the number of variables are given, choose relationship b/w only two variables.

04. **Multiple Correlation :**

When three (or) more variables are studied i.e., n variables are given, relationship b/w more than two variables.

05. **Partial Correlation :**

In this case, we study the relation between three (or) more variables and not all.

06. **Total multiple Correlation:**

In this case, we study the relationship between three (or) more variables it may be all also.

07. **Linear correlation:**

If the amount of changes in one variable a constant ratio to the amount of changes in other variable then the correlation is said to be linear.

Ex:

x	10	20	30	40	50
y	2	4	6	8	10

08. **Non-Linear correlation:**

If the amount of change in one variable does not constant ratio to the amount of change the other variable then the correlation is said to be non-linear.

Ex:-

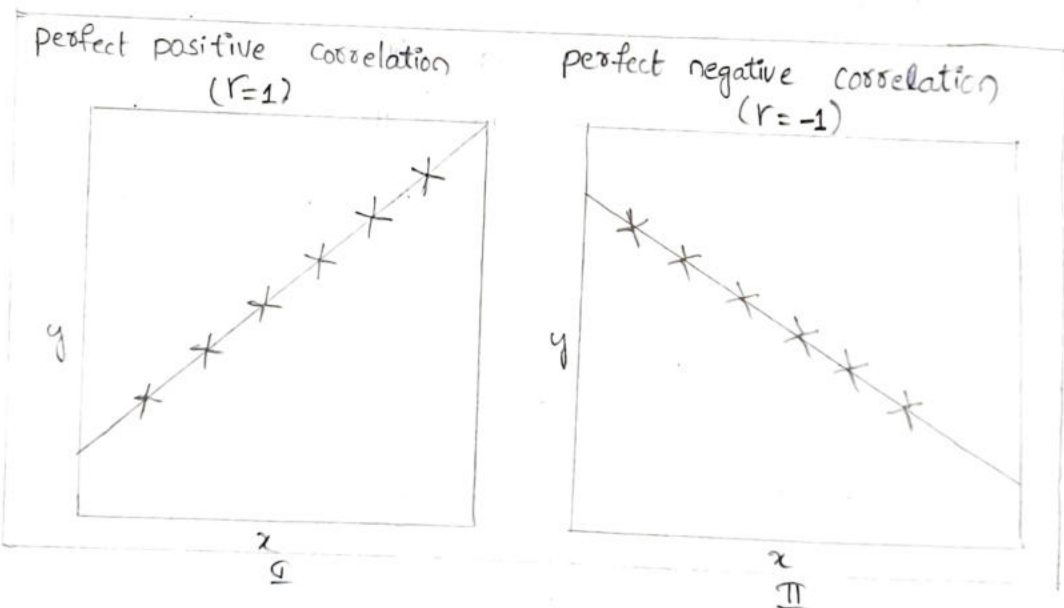
x	10	20	30	40	50
y	2	3	5	10	11

Methods of Studying correlation :

The various methods of ascertaining whether two variables are correlated (or) not are:

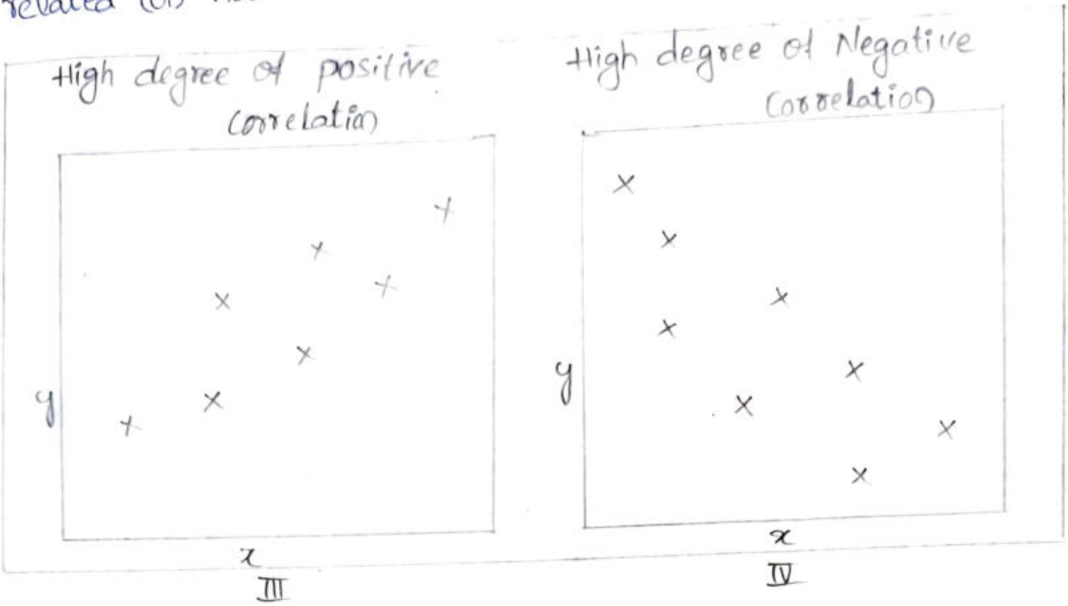
01. Scatter Diagram method.
02. Graphic method
03. Karl pearson's co-efficient of correlation
04. Concurrent Deviation method.
05. Method of Least Squares.

of these, The first two are based on the knowledge of diagrams and graphs. whereas the others are the mathematical methods. Each of these methods shall be discussed in detail in the following pages.



01. Scatter Diagram Method:

The Simplest device ascertaining whether two variables are related is to prepare a dot chart called Scatter diagram. When this method is used the given data are plotted on a graph paper (or) Simply Scatter plot in the form of dots. i.e., for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the Scatter of the Various points we can form an idea as to whether the Variables are - related (or) not.



The greater Scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line falling from the lower left-hand corner to the upper right-hand Corner. Correlation is said to be perfectly positive ($r=+1$). On the other hand, If all the points are lying on a Straight line rising from the upper left-hand corner to the lower right-hand corner of the diagram. Correlation is said to be perfectly negative ($r=-1$). If the plotted points fall in a narrow band there would be a high degree of Correlation between the Variables. Correlation shall be positive. If the points show a rising tendency from the lower left-hand Corner to the upper right-hand corner (diagram-III) and negative if the points show a declining tendency from the upper left-hand Corner to the lower right-hand Corner Of the diagram (diagram-IV). On the other hand, If the points are widely scattered over the diagram it indicates very little relationship between the Variables - Correlation shall be positive if the points are rising from the lower left-hand Corner to the upper right-hand corner (diagram(V)) and negative if the points are running from the upper left-hand side to the lower right-hand side of the diagram (diagram-VI). If the plotted points lie on a straight line parallel to the x-axis (or) in a haphazard manner. It shows absence of any relationship between the variables. ($r=0$) as shown in diagram VII.

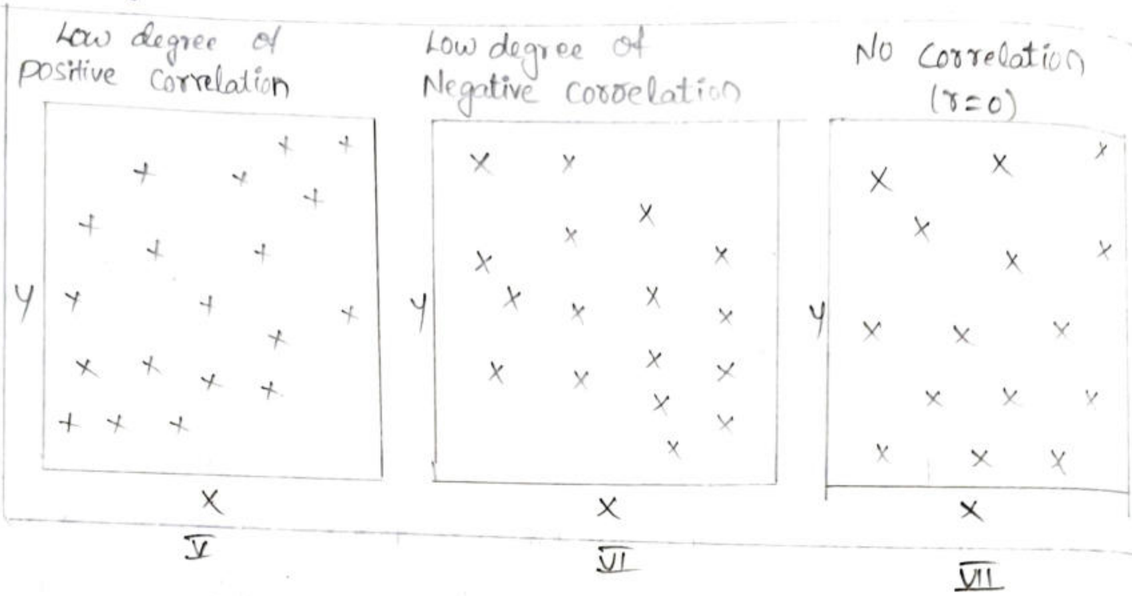
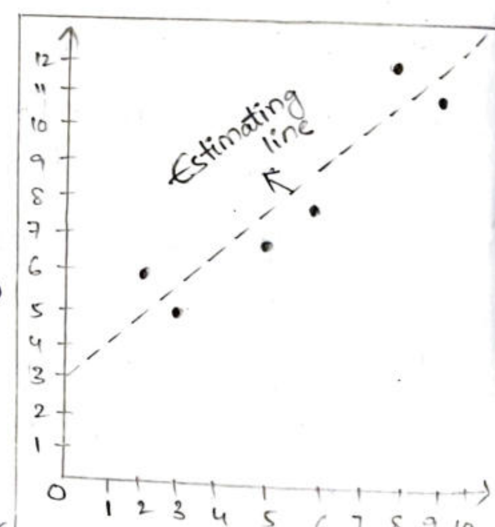


Illustration-1: Given the following pairs of values of the variables X and Y:

X:	2	3	5	5	8	9
Y:	6	5	7	8	12	11

a) Make a scatter diagram.
b) Is there any correlation between the variables X and Y?
c) By graphic inspection, draw an estimating line.



Sol:- By looking at the scatter diagram we can say that the variables X and Y are correlated. Further, correlation is positive because the trend of the points is upward rising from the lower left-hand corner to the upper right-hand corner of the diagram. The diagram also indicates that the degree of relationship is higher - because the plotted points are near to the line which shows perfect relationship between the variables.

Merit and Demerits of the Method :

Merits : Following are the merits of scatter diagram method :

→ It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.

→ It is not influenced by the size of extreme items whereas most of the mathematical methods of finding correlation are influenced by extreme items.

→ Making a scatter diagram usually is the ~~st~~ first step in investigating the relationship between two variables.

→ If the variables are related, we can see what kind of line, (or) estimating equation describes this relationship.

Demerits :

By applying this method we can get an idea about the direction of correlation and also whether it is high (or) low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.

02. Graphic method :

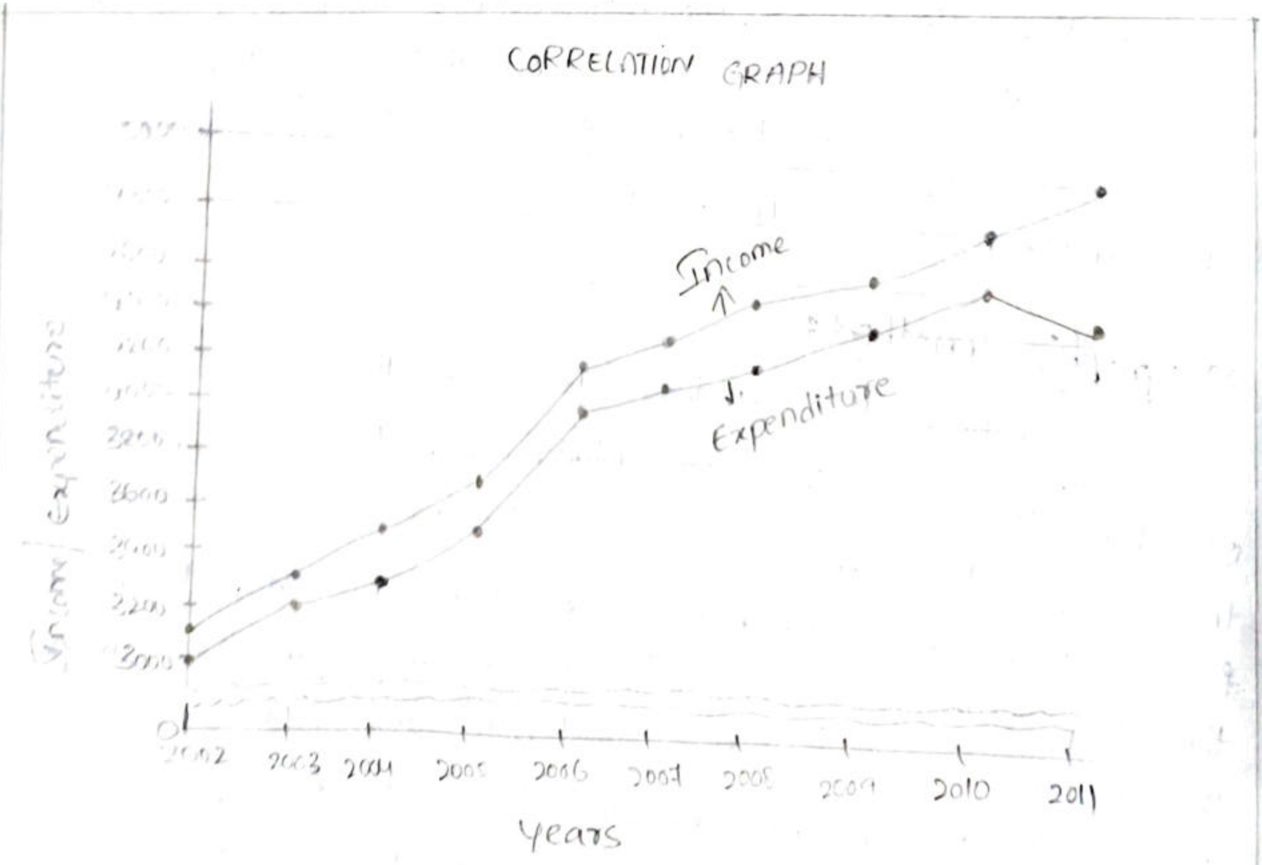
When this method is used the individual values of the two variables are plotted on the graph paper. We thus obtain two curves. One for X variable and another for Y variable. By examining the direction and closeness of the two curves so drawn we can infer whether (or) not the variables are related. If both the curves drawn on the graph are moving in the same direction (either upward (or) downward) correlation is said to be positive. On the other hand, If the curve are moving in the opposite directions correlation is said to be negative. The following example shall illustrate the method.

→ An estimating line (or) regression line is a line of average relationship. For details please see next chapter on 'Regression analysis'.

Illustration - 2 : From the following data ascertain whether the income and expenditure of the 100 workers of a factory are correlated.

Year	Average income (in ₹)	Average Expenditure (in ₹)	Year	Average income (in ₹)	Average expenditure (in ₹)
2002	3100	3000	2007	4300	4100
2003	3320	3200	2008	4500	4200
2004	3500	3350	2009	4650	4400
2005	3700	3500	2010	4800	4650
2006	4200	4000	2011	5000	4500

Sol :- The following graph shows that the variables, income and expenditure, are closely related.



This method is normally used where we are given data over a period of time, i.e., in case of time series. However, as with the scatter diagram method, in this method also we cannot get a numerical value describing the extent to which the variables are related.

03. Karl pearson's co-efficient of correlation:

Of the several mathematical methods of measuring correlation, the karl pearson's method, popularly known as pearson's co-efficient of correlation, is the most widely used in practice. The pearson co-efficient of correlation is denoted by the symbol r. It is one of the very few symbols that are used universally for describing the degree of correlation between two series. The formula for computing pearsonian r is:

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

where, $x = (X - \bar{X})$

$y = (Y - \bar{Y})$

σ_x = Standard deviation of series X.

σ_y = Standard deviation of series Y.

N = Number of pairs of observations

r = The (product moment) correlation co-efficient

This method is to be applied only where deviations of items are taken from actual mean and not from assumed mean.

The value of coefficient of correlation as obtained by the above formula shall always lie between 1 & when $r = +1$, it mean there is perfect positive correlation between the variables. when $r = -1$, It means there is perfect negative correlation between variables. when $r = 0$, it means there is no relationship between the two variables. However, in practice such values of r as +1, -1 and 0 are rare. we normally get values which lie between

This method is normally used where we are given data over a period of time, i.e., in case of time series. However, as with the scatter diagram method, in this method also we cannot get a numerical value describing the extent to which the variables are related.

03. Karl pearson's co-efficient of correlation:

Of the several mathematical methods of measuring correlation, the karl pearson's method, popularly known as pearson's co-efficient of correlation, is the most widely used in practice. The pearson co-efficient of correlation is denoted by the symbol r. It is one of the very few symbols that are used universally for describing the degree of correlation between two series. The formula for computing pearsonian r is:

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

where, $x = (X - \bar{X})$

$y = (Y - \bar{Y})$

σ_x = Standard deviation of series X.

σ_y = Standard deviation of series Y.

N = Number of pairs of observations

r = The (product moment) correlation co-efficient.

This method is to be applied only where deviations of items are taken from actual mean and not from assumed mean.

The value of co-efficient of correlation as obtained by the above formula shall always lie between 1 & when r=+1, it means there is perfect positive correlation between the variables. when r=-1, It means there is perfect negative correlation between variables. when r = 0, it means there is no relationship between the two variables. However, in practice such values of r as +1, -1 and 0 are rare. we normally get values which lie between +1 and -1 sunch as +0.8, -0.20, etc., The co-efficient of correlation describes not only the magnitude of correlation but also its direction. Thus, + 0.8 would mean that correlation is positive because the sign of r is + and the magnitude of correlation is 0.8 similarly −0.20 means low degree of negative correlation.

The above formula for computing pearson's co-efficient of correlation can be transformed to the following form which is easier to apply.

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

where, $x = (X-\bar{X})$

$y = (Y-\bar{Y})$

It is obvious that while applying this formula we have not to calculate seperately the standard deviation of X and Y series as is required by formula (i). This simplifies greatly the task of calculating correlation co-efficient.

Steps:

→ Take the deviations of X series from the mean of X and denote these deviations by x.

→ Square these deviations and obtain the total, i.e., $\sum x^2$

→ Take the deviations of Y series from the mean of Y and denote these deviations by y.

→ Square these deviations and obtain the total, i.e., $\sum y^2$.

→ multiply the deviations of X and Y series and obtain the total, i.e., $\sum xy$.

→ Substitute the values of $\sum xy$, $\sum x^2$ and $\sum y^2$ in the above formula.

The following examples will illustrate the procedure:

Illustration - 3: Calculate Karl pearson's co-efficient of correlation from the following data and interprete the values:

→ The Co-efficient of correlation is said to be a measure of covariance between two series. The covariance of two series X and Y is written as:

$$\text{Covariance} = \frac{\Sigma xy}{N}$$

where x and y stands for deviations of X and Y series from their respective means.

In order to find out the value of correlation - co-efficient, first we calculate covariance and then in order to convert it to a relative measure we divide the covariance by the standard deviation of the two series. The ratio so obtained is called Karl pearson's co-efficient.

$$r = \frac{\Sigma xy}{N} \quad \sigma_x = \sqrt{\frac{\Sigma x^2}{N}}, \quad \sigma_y = \sqrt{\frac{\Sigma y^2}{N}}$$

$$r = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N} \times \frac{\Sigma y^2}{N}}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Rollno. of students	1	2	3	4	5
marks in accountancy(x)	48	35	17	23	47
marks in statistics(y)	45	20	40	25	45

Sol:- Let marks in Accountancy be denoted by x and marks in statistics by y.

Roll no.	X	$(X-\bar{X})$ x $(\bar{X}=34)$	x^2	Y	$(Y-\bar{Y})$ y $(\bar{Y}=35)$	y^2	xy
1	48	14	196	45	10	100	140
2	35	1	1	20	-15	225	-15
3	17	-17	289	40	5	25	-85
4	23	-11	121	25	-10	100	110
5	47	13	169	45	10	100	130
	$\Sigma X=170$	$\Sigma x=0$	$\Sigma x^2=776$	$\Sigma Y=175$	$\Sigma y=0$	$\Sigma y^2=550$	$\Sigma xy=280$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{48+35+17+23+47}{5} = \frac{170}{5} = 34$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{45+20+40+25+45}{5} = \frac{175}{5} = 35$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{280}{\sqrt{776 \times 550}} = \frac{280}{653.299} = 0.429$$

Illustration : 4 : Making use of the data summarized below, Calculate the co-efficient of correlation.

Case	A	B	C	D	E	F	G	H
X_1	10	6	9	10	12	13	11	9
X_2	9	4	6	9	11	13	8	4

Sol:- Calculation of coefficient of correlation.

Case	X_1	x_1 (X_1-10)	x_1^2	X_2	(X_2-8) x_2	x_2^2	x_1x_2
A	10	0	0	9	1	1	0
B	6	-4	16	4	-4	16	16
C	9	-1	1	6	-2	4	2
D	10	0	0	9	1	1	0
E	12	2	4	11	3	9	6
F	13	3	9	13	5	25	15
G	11	1	1	8	0	0	0
H	9	-1	1	4	-4	16	4
N=8	$\Sigma X_1=80$	$\Sigma x_1=0$	$\Sigma x_1^2=32$	$\Sigma X_2=64$	$\Sigma x_2=0$	$\Sigma x_2^2=72$	$\Sigma x_1x_2=43$

$$\bar{X} = \frac{\Sigma X_1}{N} = \frac{10+6+9+10+12+13+11+9}{8} = \frac{80}{8} = 10$$

$$\bar{X} = \frac{\Sigma X_2}{N} = \frac{64}{8} = 8$$

$$r = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \times \Sigma x_2^2}}$$

$$= \frac{43}{\sqrt{32 \times 72}} = \frac{43}{48} = 0.896$$

Illustration-5: The following table gives Indices of Industrial production of registered unemployed (in hundred thousand). Calculate the value of the co-efficient of correlation.

Year	2004	2005	2006	2007	2008	2009	2010	2011
Index of production	100	102	104	107	105	112	103	99
No. unemployed	15	12	13	11	12	12	19	26

Sol:- Calculation of Karl pearson's correlation of co-efficient.

Year	X	x (X−104)	x^2	Y	y (Y−15)	y^2	xy
2004	100	−4	16	15	0	0	0
2005	102	−2	4	12	−3	9	6
2006	104	0	0	13	−2	4	0
2007	107	+3	9	11	−4	16	−12
2008	105	1	1	12	−3	9	−3
2009	112	8	64	12	−3	9	−24
2010	103	−1	1	19	+4	16	−4
2011	99	−5	25	26	11	121	−55
	$\Sigma X = 832$	$\Sigma x = 0$	$\Sigma x^2 = 120$	$\Sigma Y = 120$	$\Sigma y = 0$	$\Sigma y^2 = 184$	$\Sigma xy = -92$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{832}{8} = 104, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{120}{8} = 15$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-92}{\sqrt{120 \times 184}} = \frac{-92}{148.593} = -0.619$$

Direct method of finding out correlation Co-efficient:

Correlation Co-efficient can also be calculated without taking deviations of items either from actual mean (or) assumed mean. i.e., actual x and y values.

The formula in such a case is:

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

The formula would give the same answer as we get when deviations of items are taken from actual mean (or) assumed mean. The following example shall illustrate the point.

Illustration - 6: Calculate co-efficient of correlation from the data given below by the direct method. i.e., without taking the deviations of items from actual (or) assumed mean.

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

Sol:- Calculation of co-efficient of correlation (Direct method)

x	x^2	y	y^2	xy
9	81	15	225	135
8	64	16	256	128
7	49	14	196	98
6	36	13	169	78
5	25	11	121	55
4	16	12	144	48
3	9	10	100	30
2	4	8	64	16
1	1	9	81	9
$\sum x = 45$	$\sum x^2 = 285$	$\sum y = 108$	$\sum y^2 = 1,356$	$\sum xy = 597$

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

$N = 9$
$\sum xy = 597$
$\sum x = 45$
$\sum y = 108$
$\sum x^2 = 285$
$\sum y^2 = 1356$

$$= \frac{9\times597 - (45\times108)}{\sqrt{9\times285 - (45)^2}\sqrt{9\times1356 - (108)^2}}$$

$$= \frac{5373 - 4860}{\sqrt{2565 - 2025}\sqrt{12204 - 11664}}$$

$$= \frac{513}{\sqrt{540}\times\sqrt{540}} = \frac{513}{23.23\times23.23}$$

$$= \frac{513}{540} = 0.95$$

Calculation of correlation co-efficient when change of scale and origin is made.

Since r is a pure number, shifting the origin and changing the scale of series does not affect its values.

Illustration-7: Calculate co-efficient of correlation from the following data.

X	100	200	300	400	500	600	700
Y	30	50	60	80	100	110	130

Sol:- To simplify calculation let every value of X be divided by 100 and every value of Y by 10 and denote these series by X' and Y'.

X	$X' = \left(\frac{X}{100}\right)$	x $(X'-4)$	x^2	Y	$Y' = \left(\frac{Y}{10}\right)$	y $(Y'-8)$	y^2	xy
100	1	−3	9	30	3	−5	25	15
200	2	−2	4	50	5	−3	9	6
300	3	−1	1	60	6	−2	4	2
400	4	0	0	80	8	0	0	0
500	5	+1	1	100	10	+2	4	2
600	6	+2	4	110	11	+3	9	6
700	7	+3	9	130	13	+5	25	15
	$\sum X' = 28$	$\sum x = 0$	$\sum x^2 = 28$		$\sum Y' = 56$	$\sum y = 0$	$\sum y^2 = 76$	$\sum xy = 46$

$$\bar{X} = \frac{\Sigma X'}{N} = \frac{28}{7} = 4, \quad \bar{Y} = \frac{\Sigma Y'}{N} = \frac{56}{7} = 8$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{46}{\sqrt{28 \times 76}} = 0.997$$

When Deviations are taken from an Assumed Mean:

When actual means are in fractions, say the actual means of X and Y series are 20.167 and 29.23, the calculation of correlation by the method discussed above would involve too many calculations and would take a lot of time. In such a case, we make use of the assumed mean method for finding out correlation. When deviations are taken from an assumed mean the following formula is applicable.

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x) \times (\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

where, d_x refers to deviations of X series from an assumed mean, i.e., $(X - \bar{X})$

Similarly, d_y refers to deviations of Y series from an assumed mean i.e., $(Y - \bar{Y})$.

$\Sigma d_x d_y$ = Sum of the product of the deviations of X and Y series from their assumed means.

Σd_x^2 = Sum of the squares of the deviations of X series from an assumed mean.

Σd_y^2 = Sum of squares of the deviations of Y series from an assumed mean.

Σd_x = Sum of the deviations of X series from an assumed mean

Σd_y = Sum of the deviations of Y series from an assumed mean.

It may be pointed out that there are many variations of the above formula. For example, the above formula may be written as:

$$r = \frac{N\Sigma d_x d_y - \{(\Sigma d_x) \times (\Sigma d_y)\}}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

But the formula given above is the easiest to apply.

Steps:

→ Take the deviations of X series from an assumed mean and denote these deviations by d_x and obtain the total, i.e., Σd_x.

→ Take the deviations of Y series from an assumed mean and denote these deviations by d_y and obtain the total, i.e., Σd_y.

→ Square d_x and obtain the total Σd_x^2.

→ Square d_y and obtain the total Σd_y^2.

→ Multiply d_x and d_y and obtain the total $\Sigma d_x d_y$.

→ Substitute the value of $\Sigma d_x d_y$, Σd_x, Σd_y, Σd_x^2 and Σd_y^2 in the formula given above.

Assumption of the pearsonian co-efficient:

Karl pearson's co-efficient of correlation is based on the following assumptions:

→ There is linear relationship between the variables, i.e., when the two variables are plotted on a scatter diagram a straight line will be formed by the points so plotted.

→ The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.

→ There is a cause and effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, i.e., If the variables are independent, there cannot be any correlation. For example, there is no relationship between income and height because the forces that affect these variables are not common.

Merits and limitations of the pearsonian Co-efficient:

Amongst the mathematical methods used for measuring the degree of relationship, Karl pearson's method is most popular. The correlation co-efficient summerizes in one figure not only the degree of correlation but also the direction, i.e., whether correlation is positive (or) Negative.

However, the utility of this co-efficient depends in part on a wide knowledge of the meaning of this 'yardstick' together with its limitations. The chief limitations of the method are:

→ The correlation co-efficient always assumes linear relationship regardless of the fact whether that assumptions is correct (or) not.

→ Great care must be exercised in interpreting the value of this co-efficient as very often the co-efficient is misinterpreted.

→ The value of the co-efficient is unduly affected by the extreme items.

→ As compared with other methods this method takes more time to compute the value of correlation co-efficient.

Interpreting co-efficient of correlation:

The co-efficient of correlation measures the degree of relationship between two sets of figures. As the reliability of estimates depends upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value of co-efficient of correlation, otherwise fallacious conclusions can be drawn.

Unfortunately, the interpretation of the co-efficient of correlation depends very much on experience. The full significance of r will only be ~~(graped)~~ grasped after working out a number of correlation problems and seeing the kinds of that give rise to various values of r. The investigator must know his data throughly in order to avoid errors of interpretation and emphasis. He must be familiar or become familiar with all the relationships and theory which bear upon the data and should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. However, the following general rules are given which would help in interpreting the value of r.

→ when $r = +1$. it means there is perfect positive relationship between the variables.

→ when $r = -1$. it means there is perfect negative relationship between the variables.

→ when $r = 0$. it means there is no relationship between the variables. i.e., the variables are uncorrelated.

→ The closer r is to +1 or −1. the closer the relationship between the variables and the closer r is to 0. the less close the relation.

Rank correlation co-efficient :

The karl pearson's method is based on the assumptions that the population being studied is normally distributed. when it is known that the population is not normal (or) when the shape of the distribution is not known. there is need for a measure of correlation that involves no assumption about the parameter of the population.

It is posible to avoid making any assumptions about the populations being studied by ranking the observations according to size and basing the calculations on the ranks rather than upon the original observations. It does not matter which way the items are ranked, item number one may be the largest (or) it may be the smallest. Using ranks rather than actual observations gives the Co-efficient of rank Correlation.

→ For proof, please refer to next chapter on 'Regression Analysis'.

This method of finding out covariability (or) the lack of it between two variables was developed by the British psychologist charles Edward Spearman in 1904. This measure is especially useful when quantitative measure for certain factors (such as in the evaluation of leadership ability (or) the judgement of female beauty) cannot be fixed, but the individual in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group. Spearman's rank correlation co-efficient is defined as:

$$R = 1 - \frac{6 \Sigma D^2}{N(N^2-1)} \quad (\text{or}) \quad 1 - \frac{6 \Sigma D^2}{N^3 - N}$$

where, R denotes rank co-efficient of correlation and D refers to the difference of rank between paired items in two series.

consider a set of n individuals, ranked according to two characters x and y.

Individual	$A_1 A_2 \ldots A_i \ldots A_n$
Ranking (x)	$x_1, x_2 \ldots x_i \ldots x_n$
Ranking (y)	$y_1, y_2 \ldots y_i \ldots y_n$

$$\bar{x} = \frac{1}{N} \Sigma x_i = \frac{1}{x}[1+2+3 \ldots +N] = \frac{N+1}{2}$$

$$\bar{y} = \frac{1}{N} \Sigma y_i = \frac{1}{y}[1+2+3 \ldots +N] = \frac{N+1}{2}$$

$$\bar{x} = \bar{y}$$

$$\sigma_x^2 = \frac{N^2-1}{12} = \sigma_y^2 \text{ or } r = \frac{\mu_{11}}{\sigma_x \sigma_y}$$

Now if di stands for the difference in ranks of ith individual. we have

$$d_i = x_i - y_i = (x_i - x) - (y_i - y)$$
$$= x_i' - y_i'$$

where x_i' and y_i' are deviations of x_i and y_i for the mean $\bar{x}$ & $\bar{y}$

$$\sum d_i^2 = \sum (x_i' - y_i')^2$$
$$= \sum x_i'^2 + \sum y_i'^2 - 2\sum x_i' y_i'$$
$$\sum d_i^2 = \frac{1}{N}\sum x_i'^2 + \frac{1}{N}\sum y_i'^2 - \frac{2\sum x_i' y_i'}{N}$$
$$= \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$
$$= 2\sigma_x^2 - 2r\sigma_x^2$$
$$= 2\sigma_x^2(1-r)$$
$$\frac{1}{N} = \sum d_i^2 = 2\sigma_x^2(1-r)$$
$$(1-r) = \frac{1}{N}\frac{\sum d_i^2}{2\sigma_x^2}$$
$$(1-r) = \frac{1}{N}\frac{\sum d_i^2}{N^2\sigma_x^2} = \frac{1}{N}\frac{6\sum d_i^2}{(N^2-1)}$$
$$R = 1 - \frac{6\sum d^2}{N^3 - N}$$

The value of this co-efficient interpreted in the same way as Karl pearson's correlation co-efficient. Ranges between +1 and -1. when r_2 is +1 there is complete agreement in the order of the ranks and the ranks are in the same direction. when r_2 is -1 there is complete agreement in the order of the ranks and they are in opposite directions. This shall be clear from the following.

R_1	R_2	D (R_1-R_2)	D^2	R_1	R_2	D (R_1-R_2)	D^2
1	1	0	0	1	3	-2	4
2	2	0	0	2	2	0	0
3	3	0	0	3	1	2	4
			$\Sigma D^2=0$				$\Sigma D^2=8$

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$

$$= 1 - \frac{6\times 0}{3^3-2} = 1-0 = 1$$

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$

$$= 1 - \frac{6\times 8}{3^3-3} = 1-2 = -1$$

Features of Spearman's correlation Co-efficient:

→ The Sum of the differences of ranks between two variables shall be zero. Symbolically. $\Sigma d = 0$,

→ Spearman's correlation coefficient is distribution-free (or) non-parametric because no strict assumptions are made about the form of population from which sample observations are drawn.

→ The Spearman's correlation co-efficient is nothing but karl pearson's correlation co-efficient between the ranks. Hence, it can be interpreted in the same manner as pearsonian correlation coefficient.

In rank correlation we may have two types of problems:

→ where ranks are given.

→ where ranks are not given.

Where ranks are given:

Where actual ranks are given to us the steps required for computing rank correlation are:

⟨i⟩ Take the difference of the two ranks i.e., $(R_1 - R_2)$ and denote these differences by D.

⟨ii⟩ Square these differences and obtain the total ΣD^2.

⟨iii⟩ Apply the formula $R = 1 - \frac{6\Sigma D^2}{N^3 - N}$

problem: ⑧ The ranking of 10 students in two subjects A and B are as follows.

A R_1	B R_2	A R_1	B R_2
6	3	4	6
5	8	9	10
3	4	7	7
10	9	8	5
2	1	1	2

Calculate rank correlation co-efficient.

Sol:- Calculation of rank correlation co-efficient

R_1	R_2	D (R_1-R_2)	D^2
6	3	3	9
5	8	-3	9
3	4	-1	1
10	9	1	1
2	1	1	1
4	6	-2	4
9	10	-1	1
7	7	0	0
8	5	3	9
1	2	-1	1
N=10		$\Sigma D = 0$	$\Sigma D^2 = 36$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 36}{10^3 - 10}$$

$$= 1 - \frac{216}{1000 - 10}$$

$$= 1 - \frac{216}{990}$$

$$= 1 - 0.218$$

$$R = 0.782$$

problem : (9) Two ladies were asked to rank 7 different types of lipsticks. The ranks given by them are as follo[illegible]

Lipsticks	A	B	C	D	E	F	G
Neelu (x)	2	1	4	3	5	7	6
Neena (y)	1	3	2	4	5	6	7

Calculate Spearman's rank correlation co-efficient.

X (R_1)	Y (R_2)	D ($R_1 - R_2$)	D^2
2	1	1	1
1	3	-2	4
4	2	2	4
3	4	-1	1
5	5	0	0
7	6	1	1
6	7	-1	1
N = 7		$\Sigma D = 0$	$\Sigma D^2 = 12$

$$R = 1 - \frac{6 \Sigma D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 12}{7^3 - 7}$$

$$= 1 - \frac{72}{343 - 7}$$

$$= 1 - \frac{72}{336}$$

$$= 1 - 0.214$$

$$\therefore R = 0.786$$

problem : (10) Ten competitors in a beauty contest are ranked by 3 judges in the following order.

1^{st} Judge	1	6	5	10	3	2	4	9	7	8
2^{nd} Judge	3	5	8	4	7	10	2	1	6	9
3^{rd} Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation co-efficient to determine which pair of judges has the nearest approach to common testes in beauty.

Sol :- In order to find out which pair of judges has the nearest approach to common tastes in beauty. we compare rank correlation between the judgments of.

i) 1st judge and 2nd judge
ii) 2nd judge and 3rd judge
iii 1st judge and 3rd judge

1st judge (R_1)	2nd judge (R_2)	3rd judge (R_3)	D^2 $(R_1-R_2)^2$	D^2 $(R_2-R_3)^2$	D^2 $(R_1-R_3)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
N=10	N=10	N=10	$\Sigma D^2=200$	$\Sigma D^2=214$	$\Sigma D^2=60$

1st and 2nd judge:

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$
$$= 1 - \frac{6\times 200}{10^3-10}$$
$$= 1 - \frac{1200}{990}$$
$$= 1 - 1.212$$
$$\boxed{R = -0.212}$$

2nd and 3rd judge:

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$
$$= 1 - \frac{6\times 214}{10^3-10}$$
$$= 1 - \frac{1284}{990}$$
$$= 1 - 1.297$$
$$\boxed{R = -0.297}$$

1st and 3rd judge:

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$
$$= 1 - \frac{6\times 60}{10^3-10}$$
$$= 1 - \frac{360}{990}$$
$$= 1 - 0.364$$
$$\boxed{R = +0.636}$$

∴ Since co-efficient of correlation is maximum in the judgments of the first and third judges we conclude that they have the nearest approach to common tastes in beauty.

Where ranks are not given :

When we are given the actual data and not the ranks. It will be necessary to assign the ranks. Ranks can be assigned by taking either highest value as 1 (or) the lowest value as 1. But whether we start with the lowest value (or) the highest value we must follow the same method in case of both the variables.

Problem (iii) Calculate Spearman's Co-efficient of correlation between marks assigned to ten students by judges X and Y in a certain competitive test as shown below.

S.No.	1	2	3	4	5	6	7	8	9	10
marks by judge X	52	53	42	60	45	41	37	38	25	27
marks by judge Y	65	68	43	38	77	48	35	30	25	50

Sol :- First assign ranks and then calculate rank correlation co-efficient.

Calculation of Spearman's co-efficient of correlation

X	R_1	Y	R_2	D (R_1-R_2)	D^2 $(R_1-R_2)^2$
52	8	65	8	0	0
53	9	68	9	0	0
42	6	43	5	1	1
60	10	38	4	6	36
45	7	77	10	-3	9
41	5	48	6	-1	1
37	3	35	3	0	0
38	4	30	2	2	4
25	1	25	1	0	0
27	2	50	7	-5	25
N=10				$\Sigma D=0$	$\Sigma D^2=76$

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 76}{10^3 - 10}$$

$$= 1 - \frac{456}{990}$$

$$= 1 - 0.461$$

$$\boxed{R = 0.539}$$

problem: (12) Quotations of index no. of Security prices of a certain joint stock company are given below.

year	1	2	3	4	5	6	7
Debenture price(x)	97.8	99.2	98.8	98.3	98.4	96.7	97.1
share price(y)	73.2	85.8	78.9	75.8	77.2	87.2	83.8

Using rank Correlation method, determine the relationship between debenture prices and share prices.

Sol:- Calculation of Rank correlation co-efficient

X	R_1	Y	R_2	D^2 $(R_1 - R_2)^2$
97.8	3	73.2	1	4
99.2	7	85.8	6	1
98.8	6	78.9	4	4
98.3	4	75.8	2	4
98.4	5	77.2	3	4
96.7	1	87.2	7	36
97.1	2	83.8	5	9
				$\sum D^2 = 62$

$$R_P = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 62}{7^3 - 7}$$

$$= 1 - \frac{372}{343 - 7} = 1 - \frac{372}{336} = 1 - 1.107 = -0.107$$

Merits and Limitations of the Rank method:

Merits:

The merits of the Rank method can be discussed here.

→ This method is simpler to understand and easier to apply compared to the Karl pearson's method. The answers obtained by this method and the Karl pearson's method will be the same provided no value is repeated. i.e., all the items are different.

→ where the data are of a qualitative nature like honesty, efficiency, intelligence Etc., This method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and the degree of correlation established by applying this method.

→ This is the only method that can be used where we are given the ranks and not the actual data.

→ Even where actual data are given, rank method can be applied for ascertaining correlation.

Limitations:

The method is however associated with a few limitations too.

→ This method cannot be used for finding out correlation in a grouped frequency distribution.

→ Where the number of items exceeds 30 the calculations become quite tedious and require a lot of time. Therefore, this method should not be applied where N exceeds 30 unless we are given the ranks and not the actual values of the variables.

Equal ranks:

In some cases it may be found necessary to rant two (or) more individuals (or) entries as equal. In such a case it is customary to give each individuals an average rank. Thus if two individuals are ranked equal at

fifth place. They are each given the rank $\frac{5+6}{2}$.
that is 5.5 while. if three are ranked equal at fifth place. they are given the rank $\frac{5+6+7}{3}=6$. In other words, where two (or) more items are to be ranked equal. the rank assigned for purposes of calculating co-efficient of correlation is the average of the ranks which these individuals would have got had they differed slightly from each other.

where equal ranks are assigned to some entries an adjustment in the above formula for calculating the rank co-efficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}(m^3-m)$ to the value of $\sum D^2$. where, m stands for the number of items whose ranks are common. If there are more than one such group of items with common rank. This value is added as many times the number of such groups. The formula can thus be written.

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \dots\right\}}{N^3-N}$$

problem: (13) Compute spearman's rank correlation for the following observations:

Candidate	1	2	3	4	5	6	7	8
Judge X	20	22	28	23	30	30	23	24
Judge Y	28	24	24	25	26	27	32	30

Sol:-

Candidate	X	R_1	Y	R_2	D^2 $(R_1-R_2)^2$
1	20	1	28	6	25.00
2	22	2	24	1.5	0.25
3	28	6	24	1.5	20.25
4	23	3.5	25	3	0.25
5	30	7.5	26	4	12.25
6	30	7.5	27	5	6.25
7	23	3.5	32	8	20.25
8	24	5	30	7	4.00
					$\sum D^2 = 88.50$

$$R = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2 + \dots\right\}}{N^3 - N}$$

$$= 1 - \frac{6\left\{88.50 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right\}}{8^3 - 8}$$

$$= 1 - \frac{6\,[88.50 + 0.5 + 0.5 + 0.5]}{512 - 8}$$

$$= 1 - \frac{6 \times 90}{504}$$

$$= 1 - \frac{540}{504}$$

$$= 1 - 1.071 \quad 1.065$$

$$\boxed{R = -0.071} \quad -0.0654$$

When to use Rank correlation Co-efficient :

The Rank method has principal uses.

→ The initial data are in the form of ranks.

→ If N is fairly small (Say, not more than 25 (or) 30) rank method is sometimes applied to interval data as an approximation to the more time-consuming r. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30. The labour required in ranking the scores becomes greater than is justified by the anticipated saving of time through the rank formula.

04. Concurrent deviation method :

This method of studying correlation is the simplest of all the methods. The only thing that is required under this method is to find out the direction of change of X variable and Y variable. The formula applicable is:

$$r_c = \pm\sqrt{\pm\left(\frac{2c-n}{n}\right)}$$

where, r_c stands for co-efficients of correlation by the concurrent method ; c stands for the number of concurrent deviations (or) the number of positive signs obtains after multiplying D_x with D_y.

n = Number of pairs of observations compared.

problem: (14) Calculate the co-efficient of concurrent deviation from the following.

X	60	55	50	56	30	70	40	35	80	80	75
Y	65	40	35	75	63	80	35	20	80	60	50

Sol:- Calculation of co-efficient of concurrent deviation method.

	X	D_x	Y	D_y	D_xD_y
	60		65		
55-60	55	−	40	−	+
50-55	50	−	35	−	+
56-50	56	+	75	+	+
30-56	30	−	63	−	+
70-30	70	+	80	+	+
40-70	40	−	35	−	+
35-40	35	−	20	−	+
80-35	80	+	80	+	+
80-80	80	0	60	−	0
75-80	75	−	50	0 −	0 −
					$c = 8$

$$\therefore\ r_c = \pm\sqrt{\pm\frac{2c-n}{n}} = \pm\sqrt{\pm\frac{2\times 8-10}{10}} = \sqrt{\frac{6}{10}} = 0.774$$

problem: (15) Calculate co-efficient of concurrent deviation from the following data.

price	imports	price	imports
368	22	384	26
384	~~24~~21	395	24
385	24	403	29
381	20	400	28
347	22	385	27

Sol:- Calculation of co-efficient of concurrent deviation meth

price x	Direction of change of Variable x D_x	imports y	Direction of change of Variable y D_y	$D_x D_y$
368		22		
384	+	21	−	−
385	+	24	+	+
381	−	20	−	+
347	−	22	+	−
384	+	26	+	+
395	+	24	−	−
403	+	29	+	+
400	−	28	−	+
385	−	27	−	+
				$c = 6$

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

$$= \pm\sqrt{\pm\frac{2\times 6-9}{9}} = \pm\sqrt{\pm\frac{12-9}{9}} = \sqrt{\frac{3}{9}} = \sqrt{0.333}$$

$$\boxed{r_c = 0.577}$$

problem : (16) Calculate the co-efficient of correlation using the method of concurrent deviation between supply and Demand of an item for a period of 10 years as given below:

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Supply	125	160	164	174	155	170	165	162	172	175
Demand	115	125	192	190	165	174	124	127	152	169

Sol :- Calculation of correlation by concurrent Deviation method.

Year	X (Supply)	Dx	Y (Demand)	Dy	Dx Dy
2002	125		115		
2003	160	+	125	+	+
2004	164	+	192	+	+
2005	174	+	190	~~+~~ −	−
2006	155	−	165	−	+
2007	170	+	174	+	+
2008	165	−	124	−	+
2009	162	−	127	+	−
2010	172	+	152	+	+
2011	175	+	169	+	+
					C = 7

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

$$= \pm\sqrt{\pm\frac{2\times 7-9}{9}}$$

$$= \pm\sqrt{\pm\frac{14-9}{9}}$$

$$= \pm\sqrt{\pm\frac{5}{9}}$$

$$= \pm\sqrt{\pm 0.555}$$

$$\boxed{r_c = 0.745}$$

Merits and Limitations of Concurrent Deviation Method:

Merits:

The following are the basic advantages of this method.

→ It is simplest of all the methods

→ When the number of items is very large, this method may be used to form a quick idea about the degree of relationship before making use of more complicated methods.

Limitations:

This method is associated with the following limitations.

→ This method does not differentiate between small and big changes. For example, if x increases from 100 to 101 the sign will be plus and if y increases from 60 to 160. The sign will be plus. Thus, both get equal weight when they vary in the same direction.

→ The results obtained by this method are only a rough indicator of the presence (or) absense of correlation.

Regression Analysis :

Regression is the measure of the average relationship between two (or) more variables in terms of original units of the data.

It is mainly used to predict (estimate) the unknown value of one variable from the known values of other variable.

Types of regression :

1) Simple regression
2) multiple regression.

1) Simple regression :

If there are only two variables under consideration, then the regression is called simple regression.

2) multiple regression :

If there are more than two variables under consideration, then the regression is called multiple regression.

Lines of Regression (or) Linear Regression equations :

The lines which expresses the average relationship between two variables are known as regression lines.

If x and y are two variables, there exists two equations for regression lines.

put all small letters x and y

Method of Least Squares :-

→ Regression equation of y on x :- The regression equation of y on x is given by $\boxed{y = a + bx}$

To determine the values of a and b, the normal equations to be solved are.

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

→ Regression equation of X on Y: The regression equation of X on Y is given by $\boxed{X = a + bY}$

The normal equations are $\sum X = na + b\sum Y$

$$\sum XY = a\sum Y + b\sum Y^2$$

Method of Regression Co-efficient:

The regression line of y on x is given by

$$\boxed{y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})}$$

(or) $y - \bar{y} = b_{yx}(x - \bar{x})$

where b_{yx} is the regression co-efficient of y on x.

and $b_{yx} = r\dfrac{\sigma_y}{\sigma_x} = \dfrac{\sum xy}{\sum x^2}$

where, r = Correlation Coefficient

σ_y = S.D of Y Series

σ_x = S.D of X Series

$X = x_i - \bar{x}$

$Y = y_i - \bar{y}$

$\bar{x}$ = mean of x_i

$\bar{y}$ = mean of y_i

The regression line of x on y is given by

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

(or)

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where, b_{xy} is the regression co-efficient of x on y

and $b_{xy} = r\dfrac{\sigma_x}{\sigma_y} = \dfrac{\sum XY}{\sum Y^2}$

problem :- ① From the following data obtain the regression equations by the method of least squares.

x	6	2	10	4	8
y	9	11	5	8	7

Sol :→ Regression equation of y on x :

x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma x=30$	$\Sigma y=40$	$\Sigma xy=214$	$\Sigma x^2=220$	$\Sigma y^2=340$

$n=5$

Regression equation of y on x is $y = a+bx$

The normal equations are $\Sigma y = na + b\Sigma x$

$40 = 5a + 30b$ —— (1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$214 = 30a + 220b$ —— (2)

equation-(1) multiply with 6

(1) $\times 6 \Rightarrow 240 = \cancel{30a} + 180b$

(2) $\Rightarrow 214 = \cancel{30a} + 220b$

$(-) \quad (-) \quad (-)$

$26 = -40b$

$40b = -26$

$b = \frac{-26}{40}$

$\therefore \boxed{b = -0.65}$

'b' Value Substitute in equation - (1) $b = -0.65$

$$40 = 5a + 30b$$

$$40 = 5a + 30(-0.65)$$

$$40 = 5a - 19.5$$

$$40 + 19.5 = 5a$$

$$59.5 = 5a$$

$$\frac{59.5}{5} = a$$

$$\therefore \boxed{a = 11.9}$$

$\therefore$ Regression equation y on x = $\boxed{y = 11.9 - 0.65x}$

→ Regression equation x on y is $x = a + by$

The normal equations are $\Sigma x = na + b\Sigma y$

$$30 = 5a + 40b \longrightarrow (1)$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

$$214 = 40a + 340b \longrightarrow (2)$$

equation - (1) multiply with 8

$$(1) \times 8 \Rightarrow 240 = \cancel{40a} + 320b$$

$$(2) \Rightarrow \underset{(-)}{214} = \underset{(-)}{\cancel{40a}} + \underset{(-)}{340b}$$

$$26 = -20b$$

$$20b = -26$$

$$b = \frac{-26}{20}$$

$$\boxed{b = -1.3}$$

Substitute b Value in equation - 1 :

$$30 = 5a + 40b$$

$$30 = 5a + 40(-1.3)$$

$$30 = 5a - 52$$

$$30 + 52 = 5a$$

$$82 = 5a$$

$$a = \frac{82}{5}$$

$$\boxed{a = 16.4}$$

Regression equation x on y = $\boxed{x = 16.4 - 1.3y}$

The regression Equations are

$$\boxed{\begin{array}{l} y = a + bx \quad \text{is} \quad y = 11.9 - 0.65x \\ x = a + by \quad \text{is} \quad x = 16.4 - 1.3y \end{array}}$$

Problem: (2) Find equation of regression line y on x for the following data.

x	65	63	67	64	68	62	70	66	68	67
y	68	66	68	65	69	66	68	65	71	67

Also estimate y if $x = 75$.

Sol:-

x	y	xy	x^2	y^2
65	68	4420	4225	4624
63	66	4158	3969	4356
67	68	4556	4489	4624
64	65	4160	4096	4225
68	69	4692	4624	4761
62	66	4092	3844	4356
70	68	4760	4900	4624
66	65	$\cancel{4828}$ 4290	4356	4225
68	71	4828	4624	5041
67	67	4489	4489	4489
$\Sigma x = 660$	$\Sigma y = 673$	$\Sigma xy = 44445$	$\Sigma x^2 = 43616$	$\Sigma y^2 = 45325$

$n = 10$

Regression equation of y on x :

$$Y = a + bX$$

The normal equations are $\Sigma Y = na + b\Sigma X$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$$673 = 10a + 660b \quad \text{--- (1)}$$

$$44445 = 660a + 43616b \quad \text{--- (2)}$$

Equation-(1) multiply with 66

$$(1) \text{ --- } \times 66 \Rightarrow \quad 44418 = \cancel{660a} + 43560b$$

$$(2) \Rightarrow \quad 44445 = \cancel{660a} + 43616b$$

$$\quad (-) \qquad (-) \qquad (-)$$

$$-27 = -56b$$

$$56b = 27$$

$$b = \frac{27}{56}$$

$$\boxed{b = 0.48}$$

Substitute b value in equation-(1):

$$673 = 10a + 660b$$

$$673 = 10a + 660(0.48)$$

$$673 = 10a + 316.8$$

$$673 - 316.8 = 10a$$

$$356.2 = 10a$$

$$a = \frac{356.2}{10}$$

$$\boxed{a = 35.62}$$

Regression equation of y on x is $Y = a + bX$

$$\boxed{Y = 35.62 + 0.48X}$$

if $x = 75$,

$$y = 35.62 + 0.48(75)$$

$$= 35.62 + 36$$

$$\boxed{\therefore y = 71.62}$$

Regression equation of x on y :-

$$x = a + by$$

The normal equations are $\sum x = na + b\sum y$

$$660 = 10a + 673b \quad \text{——(1)}$$

$$\sum xy = a\sum y + b\sum y^2$$

$$44445 = 673a + 45325b \quad \text{——(2)}$$

equation -(1) multiply with 67.3,

$$(1) \times 67.3 \Rightarrow \quad 44418 = \cancel{673a} + 45292.9b$$

$$(2) \quad \Rightarrow \quad \underline{\overset{(-)}{44445} = \overset{(-)}{\cancel{673a}} \overset{(-)}{+} 45325b}$$

$$-27 = \quad -32.1b$$

$$32.1b = 27$$

$$b = \frac{27}{32.1}$$

$$\boxed{b = 0.84}$$

Substitute b value in equation -(1):

$$660 = 10a + 673b$$

$$660 = 10a + 673(0.84)$$

$$660 = 10a + 565.32$$

$$660 - 565.32 = 10a$$

$$10a = 94.68$$

$$a = \frac{94.68}{10}$$

$$\boxed{a = 9.47}$$

Regression equation x on y is $x = a + by$

$$\boxed{x = 9.47 + 0.84y}$$

Another way of calculating Regression equation:

problem: ① Calculate the Regression deviation of items from the mean of x and y Series.

x	6	2	10	4	8
y	9	11	5	8	7

Sol:-

X	x $(X-\bar{X})$	x^2	Y	y $(Y-\bar{Y})$	y^2	xy
6	0	0	9	1	1	0
2	-4	16	11	3	9	-12
10	+4	16	5	-3	9	-12
4	-2	4	8	0	0	0
8	+2	4	7	-1	1	-2
$\Sigma X=30$	$\Sigma x=0$	$\Sigma x^2=40$	$\Sigma Y=40$	$\Sigma y=0$	$\Sigma y^2=20$	$\Sigma xy=-26$

$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$ $\qquad$ $\bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$

Regression equation of x on y:

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

Hence, $x-6 = -1.3\,(y-8)$

$$x-6 = -1.3y+10.4$$

$$x = -1.3y+10.4+6$$

$$x = -1.3y+16.4$$

$$\boxed{x = 16.4-1.3y}$$

Regression equation y on x :

$$y-\bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{40} = -0.65$$

$$y-8 = -0.65\,(x-6)$$

$$y-8 = -0.65x+3.9$$

$$y = -0.65x+3.9+8$$

$$y = -0.65x+11.9$$

$$\boxed{y = 11.9-0.65x}$$

~~problem~~:

Nill

Problem: (2) Using the following ~~for~~ data obtain two Regression equations.

X	14	19	24	21	26	22	15	20	19
Y	31	36	48	37	50	45	33	41	39

Sol:-

X	x $(X-\bar{X})$	x^2	Y	y $(Y-\bar{Y})$	y^2	xy
14	-6	36	31	-9	81	54
19	-1	1	36	-4	16	4
24	4	16	48	8	64	32
21	1	1	37	-3	9	-3
26	6	36	50	10	100	60
22	2	4	45	5	25	10
15	-5	25	33	-7	49	35
20	0	0	41	1	1	0
19	-1	1	39	-1	1	1
$\Sigma X=180$	$\Sigma x=0$	$\Sigma x^2=120$	$\Sigma Y=360$	$\Sigma y=0$	$\Sigma y^2=346$	$\Sigma xy=193$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{180}{9}=20, \qquad \bar{Y}=\frac{360}{9}=40$$

Regression equation x on y:

$$X-\bar{X}=r\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$$

$$r\frac{\sigma_x}{\sigma_y}=\frac{\Sigma xy}{\Sigma y^2}=\frac{193}{346}=0.557$$

$$X-20=0.557(Y-40)$$

$$X-20=0.557Y-22.28$$

$$X=0.557Y-22.28+20$$

$$X=0.557Y-2.28$$

$$\therefore \boxed{X=-2.28+0.557Y}$$

Regression equation y on x :

$$(Y-\bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{193}{120} = 1.608$$

$$Y-40 = 1.608(X-20)$$

$$Y-40 = 1.608X - 32.16$$

$$Y = 1.608X - 32.16+40$$

$$Y = 1.608X + 7.84$$

$$\therefore \boxed{Y = 7.84 + 1.608X}$$

problem: (3) The following data relate to the scores obtained by a sales man of a company in an intelligence test and their weekly sales in Hundred rupees

Sales man Intelligence	A	B	C	D	E	F	G	H	I
Test Score(X)	50	60	50	60	80	50	80	40	70
weekly Sales(Y)	30	60	40	50	60	30	70	50	60

a) obtain the regression equation of sales on intelligence test scores of the sales man.

b) If the intelligence test score of a sales man is 65. what would be his expected weekly sales.

Sol :-

X	x $(x-\bar{x})$	x^2	Y	y $(y-\bar{y})$	y^2	xy
50	-10	100	30	-20	400	200
60	0	0	60	10	100	0
50	-10	100	40	-10	100	100
60	0	0	50	0	0	0
80	20	200	60	10	100	200
50	-10	100	30	-20	400	200
80	20	200	70	20	400	400
40	-20	200	50	0	0	0
70	10	100	60	10	100	100
$\Sigma X = 540$	$\Sigma x = 0$	$\Sigma x^2 = 1600$	$\Sigma Y = 450$	$\Sigma y = 0$	$\Sigma y^2 = 1600$	$\Sigma xy = 1200$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{540}{9} = 60, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{450}{9} = 50$$

Regression equation x on y :

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{1200}{1600} = \cancel{1.2}\ 0.75$$

$$X - 60 = 0.75(Y - 50)$$

$$X - 60 = 1.24 - 60$$

$$X = 1.24 - 60 + 60$$

$$X = 1.24$$

$$\therefore \boxed{X = 1.24}$$

Regression equation y on x :

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{1200}{1600} = 0.75$$

$$Y - 50 = 0.75(X - 60)$$

$$Y - 50 = 0.75X - 45$$

$$y = 0.75x - 45 + 50$$

$$y = 0.75x + 5$$

$$\boxed{\cancel{y = -22 + 1.2x}}$$

$$\boxed{y = 0.75x + 5}$$

if $x = 65$, $\quad y = -22 + 1.2(65)$ $\qquad y = 0.75x + 5$

$= -22 + 78$ $\qquad = 0.75(65) + 5$

$\boxed{y = 56}$ $\qquad = 48.75 + 5$

$y = 53.75;$

problem: (4) Compute the two Regression co-efficient of equations on the basis of the following information.

	x	y
mean	40	45
standard deviation	10	9

The Karl pearson correlation co-efficient between x and y is 0.50. Also Estimate value of y for x = 48.

Sol:- Given value, $\bar{x} = 40$ $\quad \bar{y} = 45$

$\sigma_x = 10 \quad \sigma_y = 9$

$r = 0.50$

Regression equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 45 = 0.50 \times \frac{9}{10}(x - 40)$$

$$y - 45 = 0.45(x - 40)$$

$$y - 45 = 0.45x - 18$$

$$y = 0.45x - 18 + 45$$

$$y = 0.45x + 27$$

$$\therefore \boxed{y = 27 + 0.45x}$$

If $x = 48$, $y = 27 + 0.45x$
$$= 27 + 0.45 \times 48$$
$$= 27 + 21.6$$
$$\therefore \boxed{y = 48.6}$$

when $x = 48$, $y = 48.6$,

Regression equation of x on y:

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 40 = 0.50 \times \frac{10}{9} (y - 45)$$

$$x - 40 = 0.556 (y - 45)$$

$$x - 40 = 0.556y - 25.02$$

$$x = 0.556y - 25.02 + 40$$

$$x = 0.556y + 15$$

$$\boxed{x = 15 + 0.556y}$$

Distinction between Correlation and Regression :

Correlation differs from Regression in the following respects:

Basis of Distinction	Correlation	Regression
1. what measures?	Correlation measures degree and direction of relationship between the variables.	Regression measures the nature and extent of average relationship between two (or) more variables in terms of the original units of the data.
2. whether relative (or) absolute measure.	It is a relative measure showing association between variables.	It is an absolute measure of relationship.
3. whether - independent of choice of both origin and scale.	Correlation Co-efficient is independent of change of both origin and scale.	Regression co-efficient is independent of change of origin and not scale.
4. whether independent of units of measurement.	correlation co-efficient is independent of units of measurement.	Regression co-efficient is not independent of units of measurement.
5. Expression of relationship.	Expression of the relationship between the variables range from -1 to +1.	Expression of the relationship between the variables may be in any of the forms like- $y = a+bx$ $y = a+bx+cx^2$
6. whether a forecasting device?	It is not a forecasting device.	It is a forecasting device which can be used to predict the value of dependent variable from the given value of independent variable.
7. Non-Sense	There may be non-sense correlation such as weight of girls and income of boys.	There is nothing like non-sense regression.

₹ 19

Business statist

Unit - V

3

Time Series Analysis

INTRODUCTION

" A time series is a set of statistical observations arranged in chronological order."

A series of observations, on a variable, recorded after successive intervals of time is called a time series. The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, and an hour, etc. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

A time series is a sequence of measurements over time, usually obtained at equally spaced intervals

– Daily

– Monthly

– Quarterly

– Yearly.

USES OF TIME SERIES ANALYSIS

1. It helps in understanding past behaviour
2. It helps in planning future operations
3. It helps in evaluating current accomplishments
4. It facilitates comparison

OBJECTIVES OF TIME SERIES ANALYSIS

The analysis of time series implies its decomposition into various factors that affect the value of its variable in a given period. It is a quantitative and objective evaluation of the effects of various factors on the activity under consideration.

There are two main objectives of the analysis of any time series data:

(i) To study the past behaviour of data.

(ii) To make forecasts for future.

The study of past behaviour is essential because it provides us the knowledge of the effects of various forces. This can facilitate the process of anticipation of future course of events, and, thus, forecasting the value of the variable as well as planning for future.

COMPONENTS OF A TIME SERIES

Components of a time series Any time series can contain some or all of the following components:

1. Trend (T)
2. Cyclical (C)
3. Seasonal (S)
4. Irregular (I)

These components may be combined in different ways. It is usually assumed that they are multiplied or added, i.e.,

$$Y = T \times C \times S \times I$$

$$Y = T + C + S + I$$

To correct for the trend in the first case one divides the first expression by the trend (T). In the second case it is subtracted.

Trend component

The trend is the long term pattern of a time series. A trend can be positive or negative depending on whether the time series exhibits an increasing long term pattern or a decreasing long term pattern. If a time series does not show an increasing or decreasing pattern then the series is stationary in the mean.

Cyclical component

Any pattern showing an up and down movement around a given trend is identified as a cyclical pattern. The duration of a cycle depends on the type of business or industry being analyzed.

Seasonal component

Seasonality occurs when the time series exhibits regular fluctuations during the same month (or months) every year, or during the same quarter every year. For instance, retail sales peak during the month of December.

Irregular component

This component is unpredictable. Every time series has some unpredictable component that makes it a random variable. In prediction, the objective is to "model" all the components to the point that the only component that remains unexplained is the random component.

TIME SERIES MODELS

The analysis of time series consists of two major steps:

1. Identifying the various forces (influences) or factors which produce the variations in the time series, and

2. Isolating, analysing and measuring the effect of these factors separately and independently, by holding other things constant.

The purpose of models is to break a time series into its components: Trend (T), Cyclical (C), Seasonality (S), and Irregularity (I).

Time series provides a basis for forecasting. There are many models by which a time series can be analysed; two models commonly used for a time series are discussed below.

1. Multiplicative Model

This is a most widely used model which assumes that forecast (Y) is the product of the four components at a particular time period. That is, the effect of four components on the time series is interdependent.

$Y = T \times C \times S \times I$ (Multiplicative model)

The multiplicative model is appropriate in situations where the effect of S, C, and I is measured in relative sense and is not in absolute sense. The geometric mean of S, C, and I is assumed to be less than one.

For example, let the actual sales for period 20 be $Y_{20} = 423.36$. Further let, this value be broken down into its components as: let trend component (mean sales) be 400; effect of current cycle (0.90) is to depress sales by 10 per cent; seasonality of the series (1.20) boosts sales by 20 per cent. Thus besides the random fluctuation, the expected value of sales for the period is $400 \times 0.90 \times 1.20 = 432$. If the random factor depresses sales by 2 per cent in this period, then the actual sales value will be $432 \times 0.98 = 423.36$.

2. **Additive Model**

In this model, it is assumed that the effect of various components can be estimated by adding the various components of a time-series.

It is stated as:

Y=T + C + S + I (Additive model)

Here S, C, and I are absolute quantities and can have positive or negative values.

It is assumed that these four components are independent of each other. However, in real-life time series data this assumption does not hold good.

3. **Mixedd Model**

It is combination above two models.

MEASUREMENT OF TREND

The principal methods of measuring trend fall into following categories:

1. Free Hand Curve or Graphic method
2. Method of Averages
3. Method of least squares

The time series methods are concerned with taking some observed historical pattern for some variable and projecting this pattern into the future using a mathematical formula. These methods do not attempt to suggest why the variable under study will take some future value. This limitation of the time series approach is taken care by the application of a causal method. The causal method tries to identify factors which influence the variable is some way or cause it to vary in some predictable manner. The two causal methods, regression analysis and correlation analysis, have already been discussed previously. A few time series methods such as freehand curves and moving averages simply describe the given data values, while other methods such as semi-average and least squares help to identify a trend equation to describe the given data values.

1. **Free Hand Curve or Graphic method**

A freehand curve drawn smoothly through the data values is often an easy and, perhaps, adequate representation of the data. The forecast can be obtained simply by extending the trend line. A trend line fitted by the freehand method should conform to the following conditions: (i) The trend line should be smooth- a straight line or mix of long gradual curves. (ii) The sum of the vertical deviations of the observations above the trend line should equal the sum of the vertical deviations of the observations below the trend line. (iii) The sum of squares of the vertical deviations of the observations from the trend line should be as small as possible. (iv) The trend line should bisect the cycles so that area above the trend line should be equal to the area below the trend line, not only for the entire series but as much as possible for each full cycle.

Example 7.1: Fit a trend line to the following data by using the freehand method.

Year	1991	1992	1993	1994	1995	1996	1997	1998
Sales turnover : (Rs. in lakh)	80	90	92	83	94	99	92	104

Solution:
presents the
graph of
turnover
from 1991 to
Forecast can
simply by
the trend

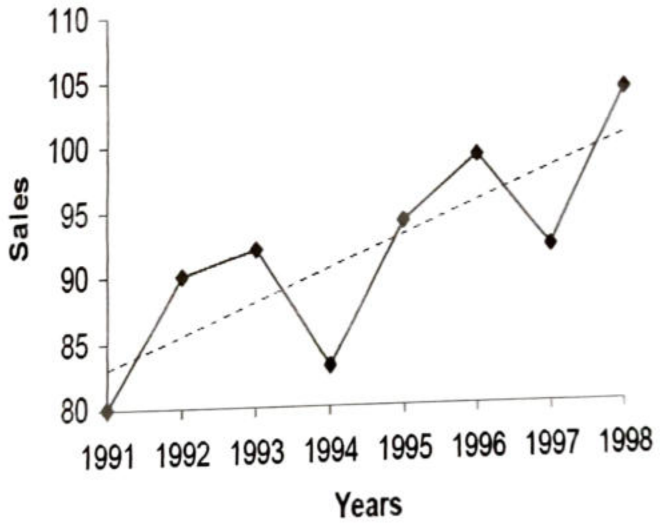


Figure 7.2
freehand
sales
(Rs. in lakh)
1998.
be obtained
extending
line.

Merits

1. It is Simplest method of measuring of trend.
2. It is very flexible
3. Hands of experience

Limitations of freehand method

(i) This method is highly subjective because the trend line depends on personal judgement and therefore what happens to be a good-fit for one individual may not be so for another.

(ii) The trend line drawn cannot have much value if it is used as a basis for predictions.

(iii) It is very time-consuming to construct a freehand trend if a careful and conscientious job is to be done.

Method of Averages

The objective of smoothing methods into smoothen out the random variations due to irregular components of the time series and thereby provide us with an overall impression of the pattern of movement in the data over time. In this section, we shall discuss three smoothing methods.

(i) Moving averages

(ii) Weighted moving averages

(iii) Semi-averages

The data requirements for the techniques to be discussed in this section are minimal and these techniques are easy to use and understand.

Semi-averages:

When this method is used, the given data is divided into two parts, preferably with the same number of years. For example, if we are given data from 1994 to 2011, *i.e.*, over a period of 18 years, the two equal parts will be each nine years, *i.e.*, from 1994 to 2002 and from 2003 to 2011. In case of odd number of years like 9, 13, 17, etc., two equal parts can be made simply by omitting the middle year. For example, if data are given for 19 years from 1993 to 2011, the two equal parts would be from 1993 to 2001 and from 2003 to 2011—the middle year 2002 will be omitted.

After the data have been divided into two parts, an average (arithmetic mean) of each part is obtained. We thus get two points. Each point is plotted at the mid-point of the class interval covered by the respective part and then the two points are joined by a straight line which gives us the required trend line. The line can be extended downwards or upwards to get intermediate values or to predict future values.

Illustration 2. Fit a trend line to the following data by the method of semi-averages :

Year	*Sales of Firm A (thousand units)*	*Year*	*Sales of Firm A (thousand units)*
2005	102	2009	108
2006	105	2010	116
2007	114	2011	112
2008	110		

Solution. Since seven years are given, the middle year shall be left out and an average of the first three years and the last three years shall be obtained. The average of the first three years is

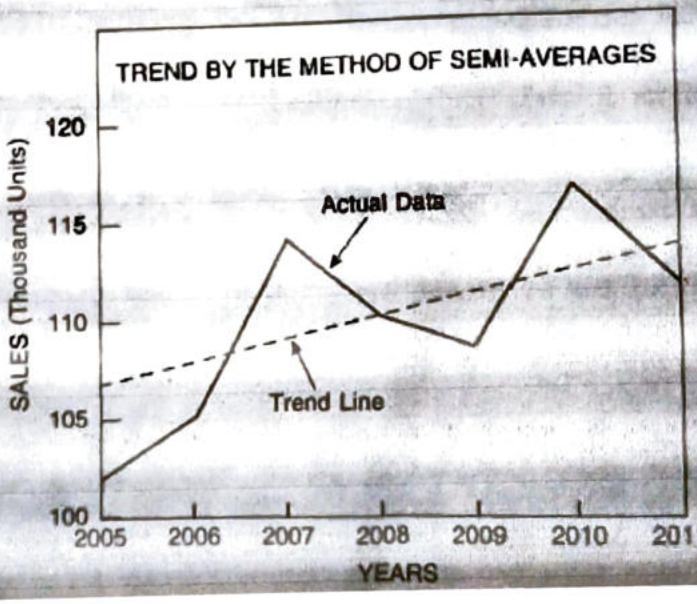
$$\frac{102 + 105 + 114}{3} = \frac{321}{3} = 107$$

and the average of the last three years is

$$\frac{108 + 116 + 112}{3} = \frac{336}{3} = 112.$$

Thus we get two points 107 and 112 which shall be plotted corresponding to their respective middle years, *i.e.*, 2006 and 2010. By joining these two points we shall obtain the required trend line. The line can be extended and can be used either for prediction or for determining intermediate values.

The actual data and the trend line are shown in the following graph :



Even Number of Years. When there are even number of years like 6, 8, 10 etc. equal parts can easily be formed and an average of each part obtained. However, when the average is to be centered there would be some problem in case the number of years is 8, 12, etc. For example, if the data relate to 2008, 2009, 2010 and 2011, which would be the middle year ? In such a case the average will be centered corresponding to 1st July 2009, i.e., middle of 2009 and 2010.

The following example shall illustrate the point :

Illustration 3. Fit a trend line by the method of semi-averages to the data given below. Estimate the sales for the year 2012. If the actual sale for that year is Rs. 520 lakh, account for the difference between the two figures.

Year	*Sales (Rs. Lakh)*		*Year*	*Sales (Rs. Lakh)*	
2004	412		2008	470	
2005	438	$\frac{1748}{4} = 437$	2009	482	$\frac{1942}{4} = 485.5$
2006	444		2010	490	
2007	454		2011	500	

Solution. The average of the first four years is 437 and that of the last four years is 485.5. These two points shall be taken corresponding to the middle periods, *i.e.*, Ist July 2005 and Ist July 2009.

The estimated sales for 2012 by projecting the semi-average trend line is Rs. 518 lakh. The actual figure given to us is Rs. 520 lakh. The difference is due to the fact that time series analysis helps us to get the best possible estimates on certain assumptions which may come out to be true or not depending upon how far these assumptions have been realised in practice.

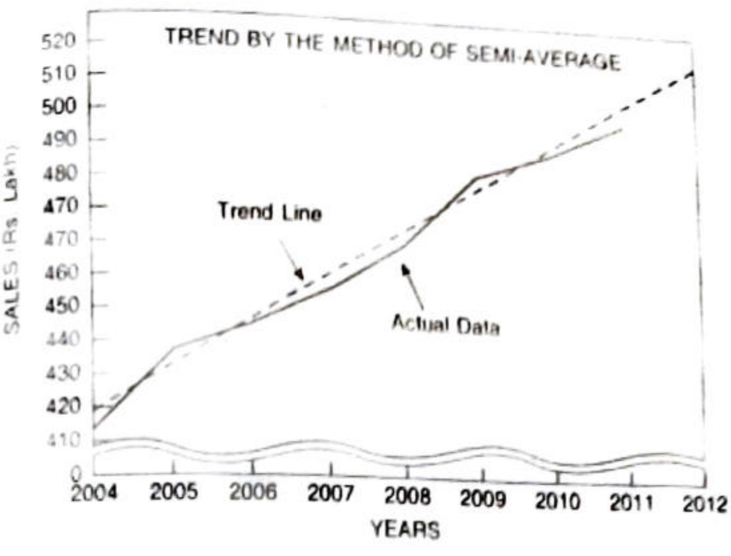


Illustration 4. The sale of a commodity in million tonnes varied from January 2011 to December 2011 in the following manner :

280	300	280	280	270	240
230	230	220	200	210	200

Fit a trend line by the method of semi-averages.

Solution.

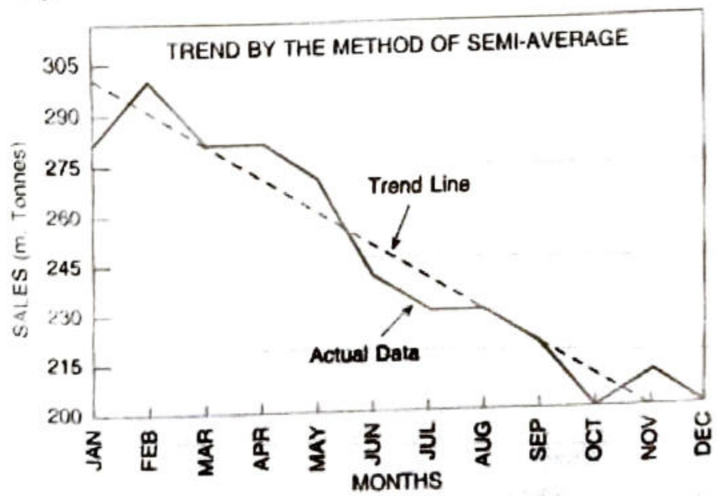
CALCULATION OF TREND VALUES BY THE METHOD OF SEMI-AVERAGES

Month	Sales in m. tonnes		Month	Sales in m. tonnes	
January	280		July	230	
February	300		August	230	
March	280	1,650 (Total) of first six months	September	220	1,290 (Total) of last six months
April	280		October	200	
May	270		November	210	
June	240		December	200	

$$\text{Average of the first half} = \frac{1650}{6} = 275 \text{ tonnes.}$$

$$\text{Average of the second half} = \frac{1290}{6} = 215 \text{ tonnes.}$$

These two figures, namely, 275 and 215, shall be plotted at the middle of their respective periods, i.e., at the middle of March-April and that of September-October, 2011. By joining these two points, we get a trend line which describes the given data.



Merits of Semi-Averages

1. The method of semi-averages is simple, easy, and quick.
2. It smooths out seasonal variations
3. It gives a better approximation to the trend because it is based on a mathematical model.

Demeris of Semi-Averages

1. It is a rough and objective method.

2. The arithmetic mean used in Semi Average is greatly affected by very large or by very small values.
3. The method of semi-averages is applicable when the trend is linear. This method is not appropriate if the trend is not linear.

Method of Moving Average

Method of Moving Averages*

When a trend is to be determined by the method of moving averages, the average value for a number of years (or month or weeks) is secured, and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the averages. The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

While applying this method, it is necessary to select a period for moving average such as 3-yearly moving average, 5-yearly moving average, 8-yearly moving average, etc. The period of moving average is to be decided in the light of the length of the cycle. Since the moving average method is most commonly applied to data which are characterised by cyclical movements, it is necessary to select a period for moving average which coincides with the length of the cycle, otherwise the cycle will not be entirely removed. The danger is more severe, the shorter the time period represented by the average. When the period of moving average and the period of the cycle do not coincide, the moving average will display a cycle which has the same period as the cycle in the data, but which has less amplitude than the cycle in the data. Often we find that the cycles in the data are not of uniform length. In such a case we should take a moving average period equal to or somewhat greater than the average period of the cycle in the data. Ordinarily, the necessary period will range between three and ten years for general business series but even longer periods are required for certain types of data.

The 3 yearly moving average shall be computed as follows :

$$\frac{a+b+c}{3},\quad \frac{b+c+d}{3},\quad \frac{c+d+e}{3},\quad \frac{d+e+f}{3}$$

and for 5 yearly moving average :

$$\frac{a+b+c+d+e}{5},\quad \frac{b+c+d+e+f}{5},\quad \frac{c+d+e+f+g}{5}$$

Illustration 5. (a) Calculate the 3-yearly moving averages of the production figures given below and draw the trend.

Year	Production (m. tonnes)	Year	Production (m. tonnes)
1997	15	2005	63
1998	21	2006	70
1999	30	2007	74
2000	36	2008	82
2001	42	2009	90
2002	46	2010	95
2003	50	2011	102
2004	56		

Solution. CALCULATION OF 3-YEARLY MOVING AVERAGES

Year	Production Y (m. tonnes)	3-yearly totals (m. tonnes)	3-yearly moving average
1997	15	—	—
1998	21	66	22.00
1999	30	87	29.00
2000	36	108	36.00
2001	42	124	41.33
2002	46	138	46.00
2003	50	152	50.67
2004	56	169	56.33
2005	63	189	63.00
2006	70	207	69.00
2007	74	226	75.33
2008	82	246	82.00
2009	90	267	89.00
2010	95	287	95.67
2011	102	—	—

(b) Construct 5-yearly moving averages of the number of students studying in a college shown below

Year	No. of students	Year	No. of students
2002	332	2007	405
2003	317	2008	410
2004	357	2009	427
2005	392	2010	405
2006	402	2011	438

Solution. CALCULATION OF 5-YEARLY MOVING AVERAGES

Year	*No. of students*	*5-yearly total*	*5-yearly moving average*
2002	332	—	—
2003	317	—	—
2004	357	1800	360.0
2005	392	1873	374.6
2006	402	1966	393.2
2007	405	2036	407.2
2008	410	2049	409.8
2009	427	2085	417.0
2010	405	—	—
2011	438	—	—

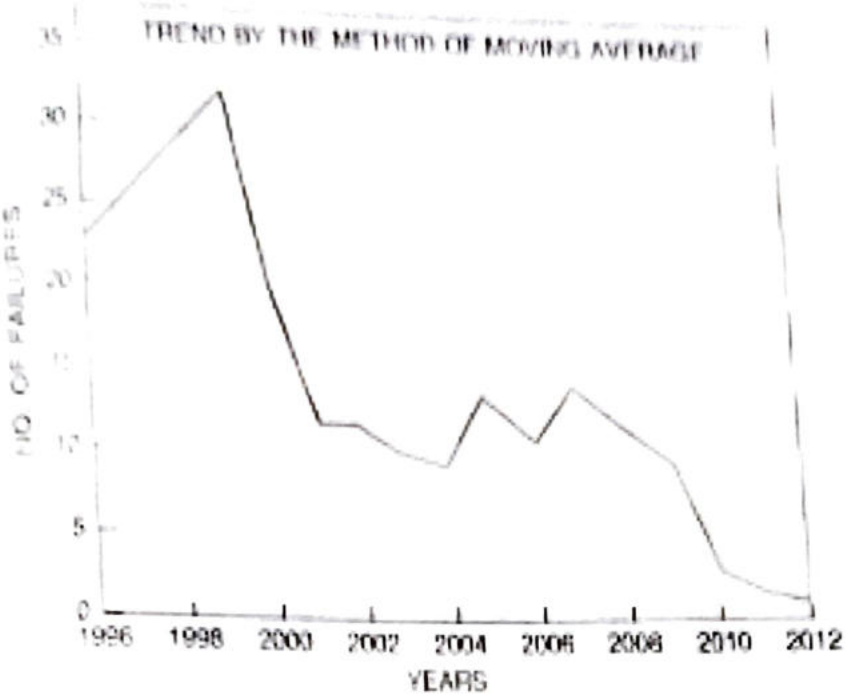
Illustration 6. Calculate 5-yearly and 7-yearly moving averages for the following data of the numbers of commercial and industrial failures in a country during 1996 to 2011 :

Year	*No. of failures*	*Year*	*No. of failures*
1996	23	2004	9
1997	26	2005	13
1998	28	2006	11
1999	32	2007	14
2000	20	2008	12
2001	12	2009	9
2002	12	2010	3
2003	10	2011	1

Also plot the actual and trend values on a graph.

Solution. CALCULATION OF 5-YEARLY AND 7-YEARLY MOVING AVERAGES

Year	*No. of failures*	*5-yearly moving total*	*5-yearly moving average*	*7-yearly moving total*	*7-yearly moving average*
1996	23		–	—	—
1997	26	—	—	—	—
1998	28	129	25.8	—	—
1999	32	118	23.6	153	21.9
2000	20	104	20.8	140	20.0
2001	12	86	17.2	123	17.6
2002	12	63	12.6	108	15.4
2003	10	56	11.2	87	12.4
2004	9	55	11.0	81	11.6
2005	13	57	11.4	81	11.6
2006	11	59	11.8	78	11.1
2007	14	59	11.8	71	10.1
2008	12	49	9.8	63	9.0
2009	9	39	7.8	—	—
2010	3	—	—	—	—
2011	1	—	—	—	—



Even Period of Moving Average. If the moving average is an even period moving average, say, four-yearly or six-yearly, the moving total and moving average which are placed at the centre of the time span from which they are computed fall between two time periods. This placement is inconvenient since the moving average so placed would not coincide with the original time period. We, therefore, synchronise moving averages and original data. This process is called *centering** and always consists of taking a two-period moving average of the moving averages.

Illustration 7. Estimate the trend values using the data given by taking a four-yearly moving average

Year	*Value*	*Year*	*Value*
1996	12	2003	100
1997	25	2004	82
1998	39	2005	65
1999	54	2006	49
2000	70	2007	34
2001	87	2008	20
2002	105	2009	7

[B.A. (H) Econ., Madras Univ., 2009]

Solution.

ESTIMATING THE TREND VALUES

Year	*Value*		4-yearly moving totals	4-yearly moving average		4-yearly moving average centered
1996	12		—	—		—
1997	25		—	—		—
		←	130	32.5		
1998	39				←	39.75
		←	188	47.0		
1999	54				←	54.75

Contd

Year	Value	4-yearly moving total	4-yearly moving average	4-yearly moving average centered
			67.5	
				70.75
[illegible]	[illegible]	[illegible]	79.0	
				84.75
2001	8[illegible]	362	90.5	
				92.00
2002	109	374	93.5	
				90.75
2003	100	352	88.0	
				81.00
2004	81	296	74.0	
				65.75
2005	65	230	57.5	
				49.75
2006	49	168	42.0	
				34.75
2007	34	110	27.5	
2008	20	—	—	—
2009	–	—	—	—

Illustration 6. Assume a four-yearly cycle and calculate the trend by the method of moving averages from the following data relating to the production of tea in India :

Year	Production (m. lbs.)	Year	Production (m. lbs.)
2002	464	2007	540
2003	515	2008	557
2004	518	2009	571
2005	467	2010	586
2006	502	2011	612

Solution. CALCULATION OF TREND BY THE MOVING AVERAGE METHOD

Year	Production (m. lbs.)	4-yearly moving totals	4-yearly moving average	4-yearly moving average centred
2002	464	—	—	—
2003	515	—	—	—
		1964	491.00	
2004	518			495.75
		2002	500.50	
2005	467			503.62
		2027	506.75	
2006	502			511.62
		2066	516.50	
2007	540			529.50
		2170	542.50	
2008	557			553.00
		2254	563.50	
2009	571			572.50
		2326	581.50	
2010	586			
2011	612	—	—	—

Illustration 9. From the following data calculate 3 yearly, 5 yearly and 7 yearly moving averages and plot the data on the graph.

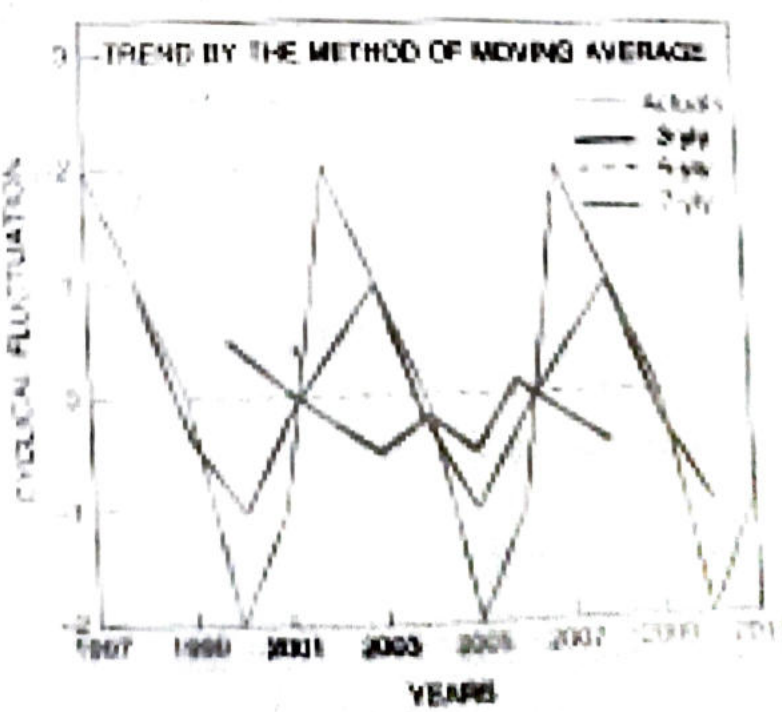
Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
Cyclical fluctuations	+2	+1	0	−2	−1	+2	+1	0	−2

Year	2006	2007	2008	2009	2010	2011
Cyclical fluctuations	−1	+2	+1	0	−2	−1

Solution

CALCULATION OF THREE-YEARLY, FIVE-YEARLY AND SEVEN-YEARLY MOVING AVERAGES

Year	Cyclical fluctuations	3 yearly moving average	5 yearly moving average	7 yearly moving average
1997	+2	—	—	—
1998	+1	+1.00	—	—
1999	0	−0.33	0	—
2000	−2	−1.00	0	−0.43
2001	−1	−0.33	0	+0.14
2002	+2	+0.67	0	+0.29
2003	+1	+1.00	0	−0.43
2004	0	−0.33	0	−0.14
2005	−2	−1.00	0	−0.43
2006	−1	−0.33	0	+0.14
2007	+2	+0.67	0	+0.29
2008	+1	+1.00	0	−0.43
2009	0	−0.33	0	—
2010	−2	−1.00	—	—
2011	−1	—	—	—



erits:

1. This method is easy to understand and easy to use because there are no mathematical complexities involved.

2. It is an objective method in the sense that anybody working on a problem with the method will get the same trend values. It is in this respect better than the free hand curve method.

3. It is a flexible method in the sense that if a few more observations are added, the entire calculations are not changed. This not with the case of semi-average method.

4. When the period of oscillatory movements is equal to the period of moving average, these movements are completely eliminated.

5. By the indirect use of this method, it is also possible to isolate seasonal, cyclical and random components.

Demerits:

1. It is not possible to calculate trend values for all the items of the series. Some information is always lost at its ends.

2. This method can determine accurate values of trend only if the oscillatory and the random fluctuations are uniform in terms of period and amplitude and the trend is, at least, approximately linear. However, these conditions are rarely met in practice. When the trend is not linear, the moving averages will not give correct values of the trend.

3. The trend values obtained by moving averages may not follow any mathematical pattern i.e. fails in setting up a functional relationship between the values of X(time) and Y(values) and thus, cannot be used for forecasting which perhaps is the main task of any time series analysis.

4. The selection of period of moving average is a difficult task and a great deal of care is needed to determine it.

5. Like arithmetic mean, the moving averages are too much affected by extreme values.