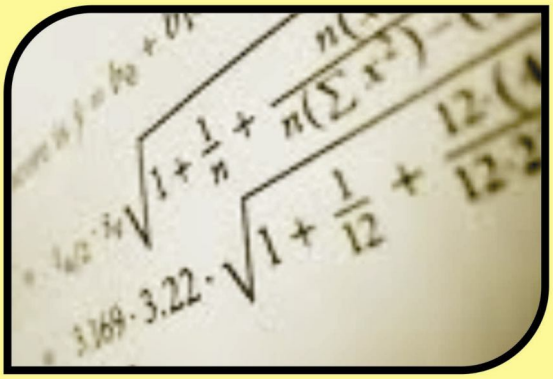
BBA-05

VARDHAMAN MAHAVEER OPEN UNIVERSITY, KOTA



BUSINESS STATISTICS

Unit - 1 Business Statistics: An Introduction

Structure of Unit:

1.0 Objectives

After completing this unit, you would be able to :

- Know the origin and growth of statistics
- Define statistics as data and as method
- Explain nature and application of statistics
- Assess the function of statistics
- Identify limitation of statistics
- Evaluate distrust and fallacies in statistics

1.1 Introduction

"When you can measure what you are speaking about and express it in numbers you know something about it, but when you can not measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind. It may be the beginning of knowledge but you have scarcely in your thought advanced to the stage of a science." **Lord Kelvin**

Facts and figures about any phenomenon-whether it relates to births, deaths, population, production, national income etc. are called 'statistics.' In this sense we can say the term statistics is considered synonymous with figures. In addition to meaning numerical facts, 'Statistics' refers to a subject. Statistics is a body of methods of obtaining and analysing data in order to base decision on them. It is branch a of scientific methods used in dealing with phenomena that can be described numerically either by counts or by measurements.

'Statistics' is being used both as a singular noun and a plural noun. As a plural noun, it stood for data while as a singular noun, it represents a method of study based on analysis and interpretation of facts. But now-a-days statistics can signify 'data' even when used as singular noun, in which case would be treated as a group noun.

Now we would be able to understand that the word statistics may mean any one of these :

(i) Numerical statement of facts or simply data,
(ii) Scientific methods to help in analysis and interpretation of data,
(iii) A measure based on sample observations.

First two of these, being more relevant to general purposes and given greater significance.

1.2 Origin and Growth of Statistics

The word statistics has originated from the Latin word 'Status', Italian word 'Statista' and German word 'Statistik' meaning a political state. The term 'Statistics' was first used in 1749 by **Gottfried Achenwall** of Germany, though the concept was in use from times back. In India, proofs about collection of data are available in the records of 'Kautilya', Magesthanese, etc. and in foreign countries also documentary evidence are there to substantiate that the data were being collected. The word 'Statist' was used by William Shakespeare, William Wordsworth and Milton etc. for a person well-versed in state administration.

Although statistics originated as a science of kings there has been a phenomenal development in the use of statistics in several varied fields. Statistics is now regarded as one of the most important tools for taking decisions in the midst of uncertainty. In fact, there is hardly any branch of science today that does not make use of statistics.

1.3 Statistics as Data

Statistics is concerned with scientific methods for collecting. Organizing, summarizing and analysing data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

There have been many definitions of the term 'statistics'. Some have defined statistics as statistical data whereas others as statistical methods. A few definitions are being discussed below: -

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

The above definition is too narrow as it confines the scope of statistics to only such facts and figures which relates to the conditions of the people in a state.

According to **Bowley** "..... numerical statements of facts in any department of inquiry placed in relation to each other."

This definition emphasises the numerical aspect of facts and extends the application of statistics to any department of enquiry in human or the physical world. It takes into consideration only the statistics which are comparable. Other primary characteristics of statistics are not explained clearly in this definition.

Prof. Horace Secrist defined statistics as "By statistics we mean aggregates of facts affected to a market extent by 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."

This definition clearly points out certain characteristics which numerical data must possess in order that they may be called statistics. It is an appropriate and exhaustive definition.

Characteristics

(i) **Aggregate of facts:** An aggregate of several facts related with an investigation is statistics. Single and isolated figures are not statistics for the simple reason that such figures are unrelated and cannot be compared and no inferences of any type can be drawn from it. For example: Production of a business in a year is not statistic, unless production of various other industries are also given. Mention of the production of different industries makes an aggregate of facts.

It should then be as stated by **Wappaus,** "aggregate of knowledge brought together for practical ends." Another statistician, **W.I. King** says. "The science of statistics is the method of judging collective national or social phenomenon from the results obtained by the analysis of an enumeration or collection of estimates."

(ii) **Numerically expressed : Prof. H. Secrist** says, "The statistical approach to a subject is numerical. Things, attributes and conditions are counted, totalled, divided, subdivided and analysed." Any facts to be called statistics must be numerically or quantitatively expressed.

Qualitative characteristics or attributes such as intelligence, beauty, rich, poor etc. cannot be included in statistics unless they are quantified by assigning certain score as a quantitative measure of assessment. For example: if we say Mohit is rich and Naman is poor, these facts will not make statistics. But the same is expressed in term of monthly income these are statistics

(iii) **Affected to a marked extent by multiplicity of causes :** Generally, facts and figures are affected to a considerable extent by number of forces operating together. For example: the phenomenon of declining sales in a business firm may be caused by a general recession in business, lack of sales promotion efforts, appearance of strong competitive forces etc. If all the relevant facts are collected and analysed it is possible to determine the factors responsible for the decline in sales. It is very difficult to study separately the effect of each of these forces. In the experimental sciences like physics and chemistry it is possible to isolate the effect of various forces on a particular event. It is proved to be a difficult task in statistical studies of phenomena which are influenced by a complex variety of factors, many of which are not measurable.

(iv) **Enumerated or Estimated:** The data may be obtained by counting or measurement or it may be estimated statistically when enumeration is not feasible or involves inordinate and high costs. Counting of numbers can be accurate to the finest degree, but the measurement of variables can be accurate only to the nearest possible extent. Estimation also gives a round about idea of the figures rather than a perfect and a precise one. If area of statistical investigation is limited, data should be collected by enumeration or counting. If area of investigation is vast or it is not possible to count, data can be collected by proper estimations also. For example: number of student in a class can be counted but number of people present in a huge crowded meeting can not be counted, it can be estimated only.

(v) **Reasonable standard of accuracy :** Data is collected only with a reasonable standard of accuracy. A high degree of accuracy as observed in accountancy or mathematics is not insisted upon in statistics because first a mass of data is involved and secondly, the process of generalisation can be achieved with a reasonable standard of accuracy only. The degree of accuracy desired largely depends upon the nature and object of the enquiry. For example: in measuring heights of persons even inches is possible whereas in measuring distance between two places can be ignored. Thus, in many statistical studies mathematical accuracy cannot be attained. However, reasonable standards of accuracy should be attained.

(vi) **Predetermined purpose:** The purpose of collecting data must be decided in advance otherwise, facts not required will be collected and that will waste labour, time and money. The definition of various terms, units of collection and measurement also help in ensuring that the data is relevant to the purpose. For example, 'data on the physical personality will be relevant for selection into military service, but it will be irrelevant for considering ability for an intellectual work.

(vii) **Systematic collection: Prof. H. Secrist** says, "stray and loose bits of quantitative information, hearsay and unrelated material gleaned here and there from indiscriminate sources having no common basis of selection, even when numerical cannot be termed as statistics." Data are to be collected in a scientific, systematic, well-planned and a properly defined way. Data collected in a haphazard manner would very likely lead to fallacious conclusions.

(viii) **Placed in relation to each other:** The main purpose of collection facts and figures is to facilitate comparative study. In other words we can say that statistics should be comparable. They are often compared period wise or region wise. For example: production of wheat in India in year 2009 can be compared with

the production of wheat in 2008 or with the production in some other country in the year 2009. But production of wheat in 2009 cannot be compared with the production of rice in 2008.

From the above discussion about the characteristics of statistics it is clear that "all the statistics are numerical statements of facts but all numerical statement of facts are not statistics."

Activity - A

Would you call following as data ? Give your answer with reason.

(i) There are three hundred student in first year commerce in college A.

_____

(ii) There are three hundred students in first year commerce, two hundred students in second year commerce and one hundred fifty students in third year commerce.

_____

(iii) The student in second year commerce are more than those in third year commerce but less than those in first year commerce.

_____

(iv) There are three hundred students in first year commerce in college A where as there are only two hundred fifty students in first year commerce in college B.

_____

(v) There are two hundred boys in first year commerce in college A where as there are one hundred girls in first year commerce in College B.

_____

1.4 Statistics as Methods

The large volume of numerical information gives rise to the need for systematic methods which can be used to organise, present, analysis and interpret the information effectively. Statistical methods are primarily developed to meet this need.

Different scholars have tried to define statistic as methods. A few of such definitions have been discussed below:

Prof. A. L. Bowley has given three definitions. At one place he says, "Statistics may be called the science of counting." But at another place he goes on to say, "Statistics may rightly be called the science of averages." At another place **Bowley** says. "Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations." All the three definitions are incomplete and focus on one aspect only in each of them. In the first definition he covers only one aspect the collection of data. Second definition also is not satisfactory because averages are only one of the devices used in statistical analysis. Third definition also is in adequate because it confines the scope of statistics only to one field, i.e. man and society, whereas statistics is applicable to every field of enquiry.

According to **A. L. Boddington,** "statistics is the science of estimates and probabilities." This definition inadequately and improperly limits statistics to only two methods estimates and probabilities, which are only a part of statistical methods.

Croxton and Cowden have given a simple and comprehensive definition of statistics. In their words, "Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data."

From a analysis of above definitions, it is clear that the methodology of statistics given by **Croxton and Cowdon** is the most scientific and realistic. It brings into light the various stages in a statistical investigation **(i)** Collection of data **(ii)** Presentation of data **(iii)** Analysis of data and **(iv)** Interpretation of data.

In simple words statistics may be called as a scientific method under which a study is made of the techniques relating to collection, organization, presentation, analysis and interpretation of numerical data.

Activity B

Whether the following problems are studied with the help of statistics ?

(i) Study of extent of honesty of a worker working in a office.

_____

(ii) Study of problem of increasing the foreign exchange reserves.

_____

(iii) A relation is to be established in the rain fall and yield.

_____

(iv) The study of effect of budget.

_____

1.5 Nature of Statistics

We can be study a subject properly only when its nature is clearly understood by us. While discussing about the nature of statistics a question comes in our mind- Is statistics a science or an art ? As a clear and definite answer to this question given by **Parson and Harlows**. He has said that by its method statistics is a science and to make use of the same is an art. Hence, we can discuss the nature of statistics under the following heads :

Statistics as a science: Science refers to a systematised body of knowledge. It studies cause and effect relationship and attempts to make generalisation in the form of scientific principles, science does not tell whether the result of a particular principle will be good or bad. We can say that science is like a lighthouse that gives light to the ships to find out their own way but does not drive them. Knowledge is said to be science when :

(1) Its study is according to some rules and dynamism.

(2) Its rules and methods are universally acceptable.

(3) It should analyse the relationship between cause and effect.

(4) It should be capable of making forecasts.

Statistics possesses all these characteristics. Statistics is studied with the help of certain rules. Statistical methods are being used widely in all fields of life and its principles are universal such as-the Theory of Probability, Law of Statistical Regularity, Law of Inertia of Large Numbers etc., the analysis of cause and effect relationship is established after collecting numerical facts and the future trends are forecasted through the techniques of interpolation, regression, analysis of time series etc.

Tippet said**,** "Statistics is a science because its methods are basically systematic and have general application." Some scholars have treated statistics as a science whereas some others do not regard it as a complete science because data are affected by multiplicity of causes and conclusions are fairly correct on an average only.

Statistics as an art : Art refers to the skill of handling facts to achieve a given objective. It is concerned with ways and means of presenting and handling data, making inferences logically and drawing relevant conclusions. Science means knowledge whereas art means action. Art tells 'how to do this." Art involves application of skill and experience for the solution of problems and gives the methods and techniques of attaining the object. The application of statistical methods requires skill and experience of the investigator. In statistics we not only learn to find various average, but also we are told which average should be used and how for as particular purpose. Hence, statistics can be called as an art.

Statistics as a science as well as an Art: According to **Tippet**, "Statistics is both a science and an art. It is a science in that its methods are basically systematic and have general application, and an art in that their successful application depends to a considerable degree on the skill and special experience of the

statistician and on his knowledge of the field of application, e.g. economics." **Harlows and Parson** described statistics as "the science and art of using the numerical facts." Hence, we conclude that statistics is not only a science, but an art also since the theoretical aspect of knowledge is an important as the practical aspect of it.

1.6 Applications of Statistics

The scope of statistics is so vast and ever-increasing that it is difficult to define it. Statistics pervades all subject matter. There is hardly any field whether it be trade, industry or commerce, economics, education, sociology. biology, botany or agriculture where statistical tools are not applicable. **Tippet** has truly said that for some subjects statistics provide basic important concepts and for some provide the methods of investigation. In this way in one form or the other the knowledge of statistics affects most of the branches.

Statistics as Business : With the gradual industrialisation and expansion of the business world, business men-find statistics as an indispensable tool. Now-a-days, the success of a particular business or industry very much depends on the accuracy and precision of statistical analysis. Before taking a new venture or for the purpose of improvement of an existing venture the Business Executives must have a large number of quantitative facts. eg. price of raw materials, price and demand of similar products, various taxes etc. All these facts are to be analysed statistically before stepping in a new enterprise or before fixing the price of a commodity. When we move forward to the globalisation of business and enter in the field of computerisation, we can not separate business with statistics.

Statistics and Economics: In the year 1890 **Prof. Marshall** the renowned economist, wrote that 'Statistics are the straw out of which I like every other economist have to make bricks." The importance of statistics in the field of economics is highlighted by this statement. Statistics has proved to be very useful in theoretical as well as practical, both the forms of economics. Economics is concerned with the production and distribution of wealth as well as with the complex institutional set-up connected with the consumption, saving and investment of income. Statistical data and statistical method are of immense help in the proper under standing of the economic problem. Statistical methods have also to be used to test and verify the hypothesis laid down by deductive logic. The deductive reasoning suggest rational human behaviour in the purchase of goods but actual observations may reveal impulsive and irrational behaviour of a large number of people under the influence of spurious advertising, aggressive salesmanship etc.

The statistical analysis plays a important role in all divisions of micro-economics and macro-economics, the law of demand and the law of elasticity of demand have been propounded on the basis of statistical analysis. Propagation of the population theory, quantity theory of money and theory of distribution etc. were possible through statistics only. Computation of national income and tax paying capacity of people is possible only with the help of relevant data. From the above facts we can say that **Tugwell** has rightly said "The science of economics is becoming statistical in its methods."

Statistics in Planning: Statistics performs a significant role in planning also. Data has to be collected on overall resources of the community including physical, financial and human resources. The information of past period is collected and used for projection into the future. Forecasting techniques are available for this purpose. Forecasting techniques based on the method of curve fitting by the principle of least squares and exponential smoothing are indispensable tools for economic and business planning. The evaluation of the progress is also an essential ingredient of planning. Statistical methods are employed both for lying down the standards and evaluating performance.

Statistics in State: Statistics is developed initially as a state craft: It was used by the rulers to assess their military and economic strength. There are references in the ancient and medieval world history that statistics was used to draw conclusions on population and land use from very early times. The registration of births and deaths in India was introduced during the Mauryan period.

In recent years the functions of the state have increased tremendously. The concept of state has changed

from that of simply maintaining law and order to that of a welfare state. Statistical data and statistical methods are of great help in promoting human welfare. Statistics help in placing suitable policies. All departments of government depend on data for their efficient working. It is impossible to fight a war successfully without data about enemy strength.

Statistics are so significant to the state that the government in most countries is the biggest collector and user of statistical data. Thus we can observe that statistics has become an essential aid to state administration and planning.

Statistics in Accounting and Auditing: The use of statistics in accounting and auditing can be admired for checking accuracy of records. Accounting information is very precise, but for decision making purpose such precision is not necessarily require and hence the statistical approximations are sought.

A very important application of statistics in accountancy is in the 'Methods of Inflation Accounting' which consist in reveling the accounting figures based on historical costs of assets by adjusting for the changes in the purchasing power of money. This is achieved through the powerful statistical tools of Price Index Numbers or the Price Deflators. Many financial and other ratios are based on statistical methods, these ratios help in averaging over a period of time. Many accounting calculations are based on statistical formulae.

In Auditing, statistical techniques of sampling are very useful in test checking. These help in determining scientifically both the size of the sample and the method of selecting the sample units. The business transactions comprised in various accounts are so large that it is practically impossible to resort to 100% examination and analysis of the records due to lack of time, money and staff. Sampling technique are used effectively to examine only a sample of the transactions and drawing inference about the whole by using the technique of estimations.

Statistics and Social Science: Statistics is also related to other social science. Statistical methods are used for the study and solution of various problems in Political science, History, Sociology, Geography, Education, Philosophy etc. such as, unemployment, literacy etc. A sociologist measures social changes and social values by statistics effectively. It collect data for knowing the housing situation of people migrated from other places. A relation can be established in pre-poll and actual results as and when elections are held in any country. Government officials collect data for evaluating the role and policies of mass media.

Statistics and Physical Science: According to **Dr. Bowley** the methods of research in statistics and physics are the same. The physical sciences, especially astronomy, geology and physics were among the fields in which statistical methods were first developed and applied. Statistical data are very much necessary for analysing the facts are deriving inferences or results in various experiments. Now-a -days physical science seem to be making increasing use of statistics. Astronomers of various countries have collected data about the movement of planets and constellations. Statistics uses in its various measurements. The position and movement of stars are being studied with the help of least squares method.

Statistics and Biological Science: It is difficult to find any scientific activity where statistical data and statistical methods are not used. In genetics different genetical analysis is done by applying techniques of theory of attributes, correlation, regression analysis etc. In diagnosing the correct disease the doctor has to rely heavily on factual data like temperature of the body, pulse rate, blood pressure etc. In judging the efficacy of a particular drug for curing a certain disease experiments have to be conducted and the success or failure would depend upon the number of people who are cured after using the drug. Statistical methods affect research in medicine and public health.

Statistics and Agriculture : The study of plant life depends upon statistics in conducting experiments about the plants, effect of temperature, type of soil etc. Experiments about different kinds of soil and fertilizers and production of different kind of crops or the growth of animals under different diets and environments are frequently designed and analysed with the help of statistical methods.

Statistical applications are not intended to be comprehensive, but they simply suggest the diversity of

applications of the underlying methods and ideas of statistics. Most of the people make use of statistics consciously or unconsciously in taking decisions **H.G. WELLS** rightly said, "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

It will not be an overstatement if we say that there is no science in recent days which works without statistics. So it is right to say "Science without statistics bear no fruit, statistics without science have no root."

1.7 Functions of Statistics

R.W. Burgers describes the usefulness of statistics as : "The fundamental gospel of statistics is to push back the domain of ignorance, prejudice, rule of thumb, arbitrary and premature decisions, traditions and dogmatism and increase the domain in which decisions are made and principles are formulated on the basis of analysed quantitative facts.

Statistical methods have taken place very fast in the social, economical, politics and every field of science. Utility of statistics is the result of important functions performed by it. **Prof Bowley** said, "The proper function of statistics is to enlarge individual experience." The important function of science of statistics are:

(1) **To Provide Numerical Facts :** One of the most important use of statistics is to present general statement in a precise and a definite form. Statement of facts conveyed in exact quantitative terms are always more convincing than the qualitatively statement. Statistics present facts in a precise and definite forms and help in proper comprehension of what is stated. Some direct facts can easily be presented in numbers like population, production, age, height, income etc. but intelligence, honesty, unemployment etc. cannot be expressed directly. These indirect facts can be changed into numerical form by statistical methods.

(2) **To Simplify Complex Facts :** The raw data is often unwieldy and complex. The complex and un-arranged facts are neither understood easily by anybody nor any result or inference can be obtained from the same. Statistics make data simple and easy to understand with the help of percentage, rates, coefficient, diagrams, graphs, tables, averages, correlation etc. **W.I. King** observes, "It is for the purpose of simplifying these unwieldy masses of facts that statistical science is useful. It reduces them to numerical totals or averages which may be abstractly handled like any other mere numbers. It draws pictures and diagrams to illustrate general tendencies and thus in many ways adopts these groups of ideas to the capacity of our intellects." If the information is presented in some systematic and condensed form it will be useful to the reader, for example: the facts about the result of student in a college for the last four years is presented only in numerical form a common person cannot draw any conclusion from these figures, but if these figures are expressed in percentage, every body can very easily understand the regular progress of the college.

(3) **To Provide Comparability :** Unless figures are compared with others of the same kind they are often devoid of any meaning. From un-comparable data we cannot interpreted accurate and correct result. Comparative study is possible by their ratios, rates, averages, percent, coefficient etc. Statistics present facts in a comparable form **A.L. Bowley** writes, "Chief practical utility of statistics is to show relative importance, that very thing an individual is likely to misjudge. Statistics are almost comparative." **Boddington** also said." The object of statistics is to enable comparison to be made between past and present results with a view to ascertaining the reasons for changes which have taken place and the effect of such changes in the future."

(4) **To Establish Relationship of Data :** Certain statistical measures such as coefficient of correlation, coefficient of association, regression etc. establish relationships between different types of data like demand and supply, supply of money and general price level, age and blindness, quantity of rainfall and electricity produced etc. We measure the change of one in comparison to the other.

(5) **To Test the Hypothesis of Other Sciences : J.M. keynes** observes, 'the function of statistics is first to suggest empirical laws which may or may not be capable of subsequent deductive explanation and secondly to supplement deductive reasoning by checking its results and submitting them to the test of

experience." The hypothesis of the theories propounded by different scientist can be tested with the help of statistical methods. The results derived by deductive method can be verified through experiments and experience. New theories can be propounded with the help of statistical methods like Angel's law of family budget and Mandel's law of heredity.

(6) **To Enlarge Individual Knowledge :** Like other science statistics also improves the knowledge and experience of individual persons. A man's view and thoughts become more clear and more firm. The knowledge of statistics assist a man in having a fair idea about any particular aspect. Many field of knowledge would have remained closed to mankind, without the efficient and useful techniques of statistical analysis. According to **Whiple,** "statistics helps in widening the horizon of knowledge and experience of a man.

(7) **To Formulate Policies and Measure Their Effects :** Statistics provide the basic material for framing suitable policies. Statistics help in formulating policies in social, economic and business field. On the basic of analysis of statistical data various government policies in the field of planning, taxation, foreign trade, social security etc. are formulated. The import- export policies of a country, prohibition policy, price policy, administrative policies, industries policies, etc. are formulated by analysing the collected data related to the matter. It really helps in taking decisions in any branch of knowledge wherever data are available. The evaluation of the effects. of any policy is also possible through statistical method. In such way, statistics proves itself as a useful tool at every stage in solving a particular problem.

(8) **To Forecast :** Almost all our activities are based on estimates about future and the judicious forecasting of future trends is a prerequisite for efficient implementation of policies. Estimates have a great importance in business decision making. Business or government plans can be made only after forecasting the trends for the future. The statistical techniques for example, extrapolation, time series, regression and forecasting are highly useful for forecasting future events.

1.8 Limitations of Statistics

The science of statistics has been profitably applied to an increasingly large number of problems concerning the administrations of business, government and in search for scientific generalisations, Unless the data are properly collected and critically interpreted there in every likely hood of drawing wrong conclusions**. According to Newsholm,** "It must be regarded as an instrument of research of great value, but having severe limitations which are not possible to overcome and as much they need our careful attention." The following are some important limitations of statistics :

(1) **Statistics studied only numerical facts** : Statistics are numerical statements of facts. Such characteristics as cannot be expressed in numbers are incapable of statistics analysis. Qualitative phenomena like honesty, intelligence, poverty etc., which cannot be expressed numerically, are not capable of direct statistical analysis. Some of such facts can be subject to a study in indirect forms, as mental level of students can be compared by the marks obtained by them, health status can be evaluated by comparing the average death rates, income of persons can be used to compare poverty or richness. But whenever the qualitative facts are measured indirectly, their accuracy and correctness is always subject to doubts and suspicious.

(2) **Statistics does not study individuals** : Statistics are aggregates of facts, so the study of an individual fact lies outside the scope of statistics. Individual facts taken separately, do not constitute statistical data and are meaningless for any statistical enquiry. It is not necessary that the conclusions of a group shall apply to individual components also. **For example :** The per capita annual income in India is Rs. 37490, but there are thousand of people who have to live emply stomach every day and few others are much wealthy. W. I King said. "Statistics from the very nature of subject cannot and will never be able to take into account individual cases."

(3) **Statistical results are true only on an average** : The conclusions obtained statistically are not universally true, they are true only under certain conditions. Statistical results useful for a general appraisal of a phenomenon

and not for substitution for any specific unit or event. Sometimes the average or trend indicated by statistics is applied to individual cases which is not proper. W.I. King writes, "Statistics largely deals with averages and their average may be made up to individual items radically different from each other."

(4) **Statistical laws are not exact :** Unlike the laws of physical and natural science, statistical laws are only approximations and not exact. On the basis of statistical analysis of problems we can talk only in term of probability and not certainty.

(5) **Statistics does not reveal the entire story :** Statistics only simplifies and helps the analysis of certain numerical facts. The interpretations of the results of statistical analysis should not be made without referring to the context. A good statistician must take cognizance of all the relevant facts before he can properly interpret the results.

(6) **Statistical methods are not the only remedy of solving problems:** For the solution of a problem statistical method is one of the methods and it is not the only approach for decision making. Result obtained from statistical methods should be confirmed by using other method also. **Croxton and cowden** have rightly said. "It must not be assumed that statistical method is the only method to use in research neither should this method be considered the best attack for every problem."

(7) Statistics is liable to be misused: The scientific methods of statistics can be understood in the most appropriate manner by the specialist only. **Yule and Kendall** write, "Statistical methods are a dangerous tool in the hands of unskilled person. These methods should be used only by those person who have a thorough and accurate knowledge of statistics, other-wise incapable and ignorant persons will arrive at misleading and wrong results."

The misuse of statistics may arise because of several reasons. If statistical conclusions are based on incomplete information one may arrive at fallacious conclusions. As **King** Says, "Statistics are like clay of which one can make a God or Devil as one pleases."

Activity C

Whether the following statement true or false, why ?

(i) The average depth of a river is 4 ft. and the average height of the members of a family is 5 ft., therefore the family can safely cross the river.

_____

(ii) Profits of company 'X' are Rs. 80,000 and profits of company 'Y' are Rs. 65,000, hence company 'X' is better than company 'Y'

_____

(iii) Average income of a village is Rs. 10,000 per month, therefore all the villagers are well paid.

_____

(iv) Births in 'P' town are more than 'Q' town, hence population in 'P' town has increased.

_____

1.9 Distrust of Statistics

Utility of statistics is universal, but this utility is not in the data only but it is also in the correct analysis and interpretation of data. People are framing a notice that anything can be proved with the help of statistics and it is true that none believes on a fact unless it is put in numerical form.

The important reason for the growing distrust is that we believe in data blindly. It is a fact that "figures do not lie but liars can figure." It can not denied that statistics are tissues of falsehood. **Desraili** has regarded statistics as a lie of lowest degree. He writes, "There are three degrees of lies - lies, damned lies and statistics."

Some of the reasons for the existence of such divergent views regarding the nature and function of statistics as follow:

(1) Figures are convincing and people are easily led to believe them.
(2) They can be manipulated in such a manner as to establish foregone conclusions.
(3) Even if correct figures are used they may be presented in such a manner that the reader is misled.

When the skilled talker writes through their forceful speeches and writings mislead the public by quoting manipulating statistical data for personal motives the public loses faith in statistics and even start condemning it.

Statistics neither prove anything nor disprove anything. Statistics is only a tool. If it properly used, it help in taking wise decision and if misused, might be disastrous. **Bowley** says, "Statistics only furnished a tool, necessary though imperfect, which is dangerous in the hands of those who do not know its use and its deficiencies." **W.I. King** also point out, "Science of statistics is the most useful servant but only of great value to those who understand its proper use."

Statistics is used in the form of double - edged sword. Data can prove any thing or what the data reveal is not so important but what they conceal is more important. According to **La Guardia**, "Statistics are like alienists ..... they will testify to either side."

Most of the fallacious ideas about the subject would be dispelled if people knew the subject and used it with care. Many of the errors would be avoided if the results are checked and verified by proper analysis.

Statistics is a very convincing tool and people use it for proving some thing which is not true and create distrust for it. **Andrew Lang** point out it as, "he sometimes used statistics as a drunkard uses a lamp post for support rather than for illumination.

1.10 Fallacies in Statistics

Some common mistakes committed in understanding and interpretations of facts are as under:

(1) **Inadequacy in collection of data :** The most important factor to be considered in statistical work is that the original collection of data is proper. If the collection of data is not adequate such as deciding the unit, scope and object of investigation, method of collection etc. are not performed properly, even the best method used in analysis and interpretation of result will not serve any useful purpose. The remark made by a judge on Indian statistics are as, "Cox, when you are a bit older you will not quote. Indian statistics with assurance. The government are very keen on amassing statistics - they collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the "Chowkidar" who just puts down when he damn pleases (quoted from "Some economic factor in Modern life", by Josiah stamp)

(2) **Inconsistency in definitions :** The errors in collection may arise because of faulty definitions which leave scope for different interpretations. The terms used in the investigation should be defined properly and clearly. If there is a slightest change in definition it may be great change in interpretation and result. Statistics collected for one purpose requires a good deal of care for use any where else.

(3) **Failure to present complete classification :** The fundamental principle of classification is that each group should be homogeneous so that its central value can truly represent the group. Lack of homogeneity often leads to fallacious results. In the absence of a complete classification, the influence of various factors may not be properly grouped.

(4) **Choice of the method:** An appropriate choice of the method is necessary to get reliable inferences from a statistical study. Only the statistician can use the statistical methods in a judicious manner. The use of these methods in contradiction to the principles would lead to wrong conclusions. By using different types of averages we can arrive at different results of the same problem. Sometimes, stating a result in terms of percentage is fallacious where absolute numbers give a better idea. Defective results can be obtained from diagrams and graphs by changing their scale.

(5) **Non-representative or inadequate data :** A basic error in statistical reasoning is to generalisation on the basis of a too small a sample which does not represent the whole. Most of the statistical conclusions

are inducted by the application of the technique of sampling. An error of induction would be generated if a generalisation may be made from insufficient or non-representative sample.

(6) **Inappropriate comparisons :** Comparisons are a part of analysis but a good deal of care is required to avoid fallacious conclusion. Comparisons should be not being between phenomena which are not comparable. Two sets of data can be compared only when they are homogeneous as regards nature, kind, scope, time, place etc. Comparison of heterogeneous data will give fallacious conclusions.

(7) **Bias on the part of investigator :** Statisticians sometimes do have a conclusion already decided upon and choose their sample to prove their conclusion. Some persons do make assumptions that what they know is not justified and hide their doubt, so statistical conclusions are also rendered incorrect because of prejudice of the investigator, because such investigator always tries that result should be in his favour.

1.11 Summary

Statistics means numerical presentation of facts. In the globalized world every branch of knowledge is directly or indirectly associated with numerical facts. Analysis of numerical facts helps one in arriving at certain conclusions. The use of world 'statistics' in the sense of data is a narrower use of the term. The broader use of terms in reference to statistical method which means study of certain methods, principles and techniques by which data are collected, analysed, interpret and draw relevant conclusions. Statistical methods are to be used as a tool for solving a variety of problems and its application is universal.

It is obvious that the statistical tools have to be used with great care then the conclusion and inferences have to be drawn with proper understanding of the situation and the object of collection, classification and analysis and purpose for which the data was gathered has always to be kept in mind. If we use statistical data carefully and cautiously we will get accurate results, but if misused the results will be misleading and fallacious.

1.12 Key Words

Data : Aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose.

Statistical Method : Procedures used in collection, organisations, summary, analysis, interpretation and presentation of data.

Business Statistics : Use of statistical methods to study, analysis and find solutions to various problems of the business.

1.13 Self Assessment Questions

Q.1 "All statistical data are numerical statement of facts, but all numerical statement of facts are not statistical data." Explain this statement, State the characteristics of statistical data.

Q. 2 Define statistics and explain the functions and importance of statistics.

Q. 3 Define statistics. Discuss its scope and limitations.

Q. 4 Comment on the following statements.
(i) There are three lies - lies, damned lies and statistics.
(ii) Figures do not lie.

Q.5 Enumerate causes of distrust of statistics.

1.14 Reference Books

1. Sancheti, D.C.; Kapur, V. K., Statistics (Theory, Methods & Applications).

2. Gupta, S.P. Statistical Methods.

3. Elhance, D.N. Elements of Statistics.

Unit - 2 Statistical Investigation

Structure of Unit:

2.0 Objectives

After Completing this unit, you will be able to :

- Assess the necessity of investigation.
- Evaluate the planning of investigation.
- Explain various stages of statistical investigation.
- Define statistical unit.
- Discuss execution of survey.
- Explain advantages and disadvantages of investigation.

2.1 Introduction

Numerical data constitutes the raw material for statistical analysis. Data can be obtained through a statistical survey also called statistical investigation. A statistical survey is nothing but a systematic search for truth. It seeks some authentic answers to a problem which is quantifiable and amenable to statistical treatment.

Statistical investigation is a part of an information gathering and learning process which is undertaken to seek meaning from and to learn more about observed phenomena as well as to inform decisions and actions. The ultimate goal of statistical investigation is to learn more about a real world situation and to expand the body of contextual knowledge.

A statistical survey may be either a general purpose survey or special purpose survey. In a general purpose survey we obtain data which are useful for several purposes. A special purpose survey is that in which data obtained are useful in analysing a particular problem only.

2.2 Meaning and Necessity of Investigation

Investigation means "to search for the knowledge". Statistical Investigation refers to the search of knowledge about some fact or problem which is provided or obtained through the use of statistical methods. Knowledge should be in form capable in quantitative measurement and analysis. Statistical investigation is gaining quantitative knowledge through statistical methods. The entire process of investigation is carried through analysing numerical data. Thus, data are the fundamental factor of statistical investigation.

Statistical Investigations are used to collect quantitative information about items in a popultion. Survey of human populations and institutions are common in political polling and government, health, social science and marketing research. An investigation may focus on opinions or factual information depending on its purpose and many investigations involve administering question to individuals. When the questions are administered by a researcher the investigation is called a structure interview or a researcher administered survey. When the questions are administered by the respondent, the survey is referred to as a questionnaire or a self administered survey.

2.3 Planning of Investigation

The Plan involves working out what we are going to do to solve the problem. Deciding what we will measure and how we will measure it. We also need to decide how we will collect and record our data.

Proper planning of a survey is of paramount importance because the quality of survey results depends considerably on the preparations made before the survey is conducted. A well designed lay out of goal makes the task of the investigator much easier. During the process of formulating the plan, he comes to know of several difficulties and for the same he can make suitable arrangements.

According to **Parten** "Only by careful planning the survey from start to finish can reliance be placed upon results." We should consider all the aspect of the problem very well, e.g., - what is the object of investigation? What information is necessary for the same ? from where and how can this information be collected etc. If the scheme is pondered over thoroughly and work is started according to plan, the success of the purpose is certain.

The matters which require careful consideration at the planning stage may be enumerated as follows:

- Specification of the purpose
- Scope of the survey
- Define the problem
- Sources of data
- Technique of data collection
- The frame
- The forms of enquiry
- Determination of statistical unit
- Degree of accuracy desired

Specification of the purpose : The objective of a statistical survey should be clearly set out before survey is conducted. This will invariably indicate the type of information which is needed and the use to which the information obtained will be put. Which statistical method to be employed will also depend upon the purpose. The object of an enquiry may be either to collect specific information relating to a problem or adequate data to test a hypothesis for a given proposition. Collection of irrelevant data without any purpose will waste resources as time, money and energy. In surveys which are likely to provide information that will be of value to different organisations or government departments, a detailed statement of the uses to which the information obtained could be put may be prepared. It will be helpful to all organisations and they might be expected to make suggestions for suitable modifications before the survey begins.

According to **Prof. W.A. Neiswanger** "Statement of objective is the basic importance because it determines the data which is to be collected."

Scope of the survey : The object of the problem would throw light on the scope of the investigation. This explains what aspects have to be covered to achieve the given objectives. Its coverage with regard to the type of information, the subject matter and geographical area. For example: an enquiry may be related to country as a whole, or one particular state or an industrial town only. As much as the area of investigation is wide the more representative are likely to be the results. However, it will depend upon the object. The expenditure to be incurred by the investigator and time devoted are also important in the determination of scope.

Three factors exert great influence on scope - The object of enquiry, availability of time and availability of resources. The investigation should be carried out with in a reasonable period of time, otherwise information collected may become out-of-date and have no use. Delay may also results in losses. The scope of an enquiry usually fixes the limits of the enquiry. A certain amount of discretion can always be exercised in this respect.

Define the problem : The problem must be defined in exact and precise term. There may be various terms in the statement of purpose and scope of work which might mean many things. These various terms should be properly defined to avoid any mistake due to use of terminology. It may be defined in accordance with the object of investigation. If the problem has not been defined very well several difficulties may arise during the process of investigation and collection of data may prove to be useless. If the problem has been generalised, it may lead to collection of erroneous data and consequently fallacious conclusion. Vagueness of the problem should be removed by attaching specific meaning to the words which are capable of double or multiple interpretations. For example: whether profitability would refer to absolute levels of profit or a rate of return on sales turnover capital employed etc. should be clearly, mentioned. The true meaning of sales turnover and capital employed will have to be specially given to avoid confusion in the mind of the persons conducting the survey.

Sources of data : After the purpose and scope have been defined, then investigator have to decide about the source of data. The data can be collected through primary or secondary sources. When the investigator collects first hand data for the purpose at hand i.e. data will have to be gathered from the original field of investigation, such data are known as primary data. If the investigator obtained the data from published or unpublished sources, the original field not to be approached afresh, such data will constitute secondary data for him. Investigator may use primary or secondary data or both type of data may be used by him in a particular investigation. It is depends upon the purpose and scope of investigation. The purpose of enquiry, the availability of time and resources play an important role in shaping the mode of enquiry.

Technique of data collection : There are two important techniques of data collection. These are **(i)** census technique and **(ii)** sample technique. A census is a complete enumeration of each and every unit of the population i.e. investigator studies every single unit in the census technique whereas in a sample only a part of the population is studied and conclusions about the entire population are drawn on that basis.

If each individual unit of the population is important or these are widely different from each other, then use of census method is appropriate and if there is uniformity among the various units, sample method is appropriate. The census method is more expensive and time consuming as compared to the sample method. The investigator must decide which technique he will use. The decision would depend upon these factors: **(i)** the availability of resources, **(ii)** the time factor, **(iii)** the degree of accuracy desired and **(iv)** nature and scope in the problem.

The frame : The term 'frame' refers to a list map or other specification of the units which constitute the available information relating to the whole designated for a particular survey scheme. The whole structure of enquiry is to a considerable extent determined by the frame. The method of survey which is suitable for a given type of material may be different in various territories because different types of frames have to be used. Detailed planning of the survey cannot be undertaken without nature and accuracy of the available frames are known. Same frames may not be suitable for different survey. There may be various types of defects existing in the available frame. Frame may be inaccurate, incomplete, subject to duplicate, inadequate or out of date. When the detailed investigation has been made, it is necessary for a survey to carry out a careful investigations of the frame that is proposed to be adopted since defects are not apparent.

Form of enquiry : Different types of investigations are suitable for the enquiry into different kinds of objects. Selection of the form of investigation depends upon the ojbect, scope, cost etc. There may be different form of investigation.

(A) **Direct or indirect -** Which type of enquiry should be made, this depends on the nature and

object of investigation. Where data are capable of direct quantitative measurement as height, weight, length etc. is known as direct survey. When the direct quantitative measurement is not possible as poverty, intelligence, efficiency is known as indirect enquiry. In the indirect enquiry, investigator has to take up certain objective measurement phenomena which reflect the qualitative phenomena and then relevant data to be collected. For example: the intelligence of students can be studied with the help of data about marks secured by them.

(B) **Official, semi-official or non-official -** When the investigation conducted by the central or state government, it is called official investigation, public interest is also involved in such type of investigation. In case of official enquiry people might be bound by law to provide the require information. When the investigation conducted by the semi-government or autonomous bodies like university, municipal corporation etc., it is called semi-official investigation. In case of semi-official enquiry people may release information on personal requests. When the investigation conducted by private bodies or individuals such as business man, political parties, newspapers etc. it is known as non- official enquiry. In such investigations, interest of the investigator is involved and no body can be forced to supply the information.

The facility available will differ according to the nature of enquiry. Official or semi official investigations are easy to conduct because of the availability of sufficient information whereas it is a difficult job to conduct non-official investigation because people are highly hesitant to disclose correct information, because they have a tendency to hide and conceal facts.

(C) **Regular or adhoc investigation-** When an investigation takes place at regular intervals over a definite period of time, it is a regular survey, for example: Reserve Bank collects the data of loan disbursed by commercial banks regularly after every 3 months. An enquiry is conducted for a specific purpose without any regularity, it is known as adhoc investigation.

(D) **Confidential or non-confidential investigation -** If the results of the enquiry are not to be disclosed or it is not in public interest to publish the result, such enquiry known as confidential investigation. On the other hand, if the scope of investigation is vast and the result of the enquiry are published in the public interest, it is known as non-confidential or open investigation. A confidential enquiry may be brought into open later on if deemed proper to do so.

(E) **Initial or Repetitive investigation -** An initial investigation is one that is carried out for the first time. In the case of initial investigation it is necessary to formulate a plan for collection of data whereas repetitive investigation is one that is conducted in continuation of previous enquiry. In this type of investigation, a plan already exists and may only need modification to it, in the view of past experience.

Determination of statistical unit : Statistical unit must be determined for collection of data. Statistical unit will be quantitative and not qualitative. The unit must be precise and clearly defined. It is necessary not only for collection of data but also for analysis and interpretation. The unit in tems of which the investigator counts or measures the variable or attributes selected for enumeration analysis and interpretation is known as a 'statistical unit'.

Degree of accuracy desired : The investigator has to decide about the degree of accuracy that he wants to attain. How much accuracy is desired, it depends on the object of investigtion. Absolute accuracy is seldom possible to be achieved in every field of statistical investigation because **(i)** statistics are based on estimates, **(ii)** tools of measurement are not always perfect, **(iii)** there may be unintentional bias on the part of the investigator, enumerator or informant. In scientific investigations, absolute accuracy may be possible but in social, economic and commercial fields neither it is feasible nor desirable. If an attempt is made to attain 100% accuracy it would not be realistic. It requires much money time and labour to attain high degrees of accuracy which is beyond the capacity of most of the investigators. A reasonable degree of accuracy must be aspired for and attained. The reasonableness of accuracy will vary from investigation to investigation according to circumstances. **Riggleman and frishbee** rightly says. "The necessary degree of accuracy in counting or measuring depends upon the practical value of the accuracy in relation to its

cost." We don't say that one should sacrifice accuracy to keep down the costs. It would ultimately depend upon the purpose, nature and scope of investigation. It is desirable that an eye be kept on the possible inaccuracies that are likely to arise due to clerical and other type of errors to that they may be eliminated altogether or reduced to the minimum.

While planning a statistical enquiry keeping in view the above point, a systematic program should be formed so that the process of investigation may continue. Some of above stated precautions are nothing more than commonsense, but are nonetheless worth nothing. If they are neglected, the result may be completely useless.

Activity A

Comment on the following from the point of view of a statistical investigator :

(i) According to a person the number of literate people in a village at a time was 500 and according to the other at a same time it was 800.

_____

(ii) The company producing 'Ponds' powder wants to bring a new quality of powder in the market.

_____

(iii) With the objective of printing the poster, data about the cost of its printing were collected.

_____

(iv) Information about number of students of schools situated in a area was collected.

_____

2.4 Statistical Unit

There is a basis of measurement which is also known as scale, for counting, weighing, measuring the various items. Before organising the task of collecting data the statistical unit or units must be clearly defined for purpose of investigation. The problem of defining the unit is not as simple as it appears to be.

Statistical units are those means of measurement in terms of which data are collected. These are analysed and presented also in the same terms on the same basis. The basis of determining the size must be clearly defined and the same definition followed throughout the investigation. For example : distance is measured in kilometers, height in centimeters, production in tonne etc., then kilometres, centimetres, tonne etc. will be the statistical unit.

According to **Prof. king,** "It is not only appropriate but necessary also that the definition of statistical unit is simple, clear and unmistakable."

2.5 Characteristics to be present in a Statistical Unit

While fixing the statistical unit for an enquiry, it is useful to keep in view the following characteristics:

(i) **Simple and clear :** Statistical unit should be well defined, simple to understand and definite so that uniformity may be maintained in calculation and people may not confused for the meaning. For example: meaning of unemployment, literacy etc. should be well defined.

(ii) **Stability and standard** : The unit must be exact and precise means, its value remains always constant or stable. It should be acceptable to all and it must be standardised one. If a value for measurement decided at the beginning of investigation, it will be remain unchanged during whole process of investigation. Unit, which will be used should be the standard unit throughout the country. The unit must be invogue and used by majority of people.

(iii) **Uniformity or homogeneity :** A statistical unit should be uniform throughout the study so there can be comparisons possible. If units are defined differently, at different stages of investigation, comparison would become difficult. They would also lead to wrong or absurd conclusions and collection of data will be useless also.

(iv) **Suitability :** The suitability of a unit should be decided keeping into consideration the object and scope of investigation. For example: if the investigation is about distance between two cities, kilometers should be selected for measurement as unit, but if the height of a man to be measured, unit should be selected as centimeters or ft.

Types of statistical units : The statistical units broadly classified under two heads :

(i) Units of collection
(ii) Units of analysis and interpretation

(i) **Units of collection :** The units used for collection of data are known as units of collection. They involve either counting or measurement. Number of people, houses, articles are unit of counting, kilometers, kilograms, rupees are unit of measurement. In the process of collection investigator may deal with either discrete, entities and events relating to them as in the case of persons, houses, number of items etc. or with measurable quantities and value units such as kilometers, ton etc. Such units can be simple, composite or complex.

Simple unit: - A simple unit is one which represents a single condition without qualifications. For example : a house, a child. In some condition such unit should be carefully defined as a dozen contain 13 pieces or 12 pieces should be clearly defined.

Composite unit - When two simple units are combined together, it forms a composite unit, A compound unit is a simple unit the comprehension of which is subject to some qualification. For example: simple unit worker may be qualified as 'skilled worker' or 'unskilled worker.' A complex unit is formed by adding to a simple unit two or more qualifications. For example: in service industries just as roadways and railways, while finding the cost of travelling and carrying load, passenger - kilometers and ton - kilometers are taken as convenient units of measurement.

(ii) **Units of analysis and interpretation :** Statistical data are collected for making comparisons. We can also make comparison reference to time or space. Units on the basis of which data are compared are known as the units of analysis and interpretation. Units of analysis and interpretation are those facilitate comparisons. We can also understand data more easily through these expressions.

Rate : Rate are used in those cases where comparisons are made between quantities of different kinds. When two quantities are to be compared and they in the form of ratios, but the denominator and numerator are different, then we express them in rates. Rates are mostly expressed as percent or per thousand.

Ratio : Ratio are used where quantities to be compared are of the same kind. When we want to expressed the relation between two homogeneous or similar quantities, such as men and women, we can express them in form of ratio. If same figures are to be placed in the form of ratio, all of them can be multiplied or divided by a common figure.

Coefficient : If the absolute figure is expressed in relation to some other connected variable or in relation to the figure on the basis of which the absolute figure has been arrived at, is known as coefficient. In simple words we can say a unit which is used for comparison between similar numerator and denominator is known as coefficient. It is also called rate per unit. We can also observe that it is such a number which on multiplication with the total gives the respective number.

2.6 Stages of Statistical Investigation

Croxton and cowden have given a very simple definition of statistics. According to them, "Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data." The definition points out four stages of statistical investigation. The followings are stages of actual operation of investigation :

Collection of data : Collection of data constitutes the first step in a statistical investigation. There are

different methods for collecting data, Utmost care must be exercised in collecting data because they form the foundation of statistical analysis. The data may be available from existing published or unpublished sources or may be collected by investigator himself. The collection of primary data is most difficult and important work of investigator. Questionnaires and schedules of different kinds are required to be prepared by him. The investigator must take into account whatever data have already been collected by others. This would save the investigator from unnecessary labour and duplication of efforts.

Organization: The organisations consist of three steps viz., editing, classification and tabulation. Collected data are edited according to the object of investigation. Secondary data are generally in organised form. Data that are collected by the investigator needs editing. It is better to remove the material not needed for investigation. Data must be edited very carefully so that the omissions, inconsistencies, and irrelevant answers may be corrected or adjusted. After editing the next step is to classify the data. By classification, collected data may be easy to grasp and comparable. The object of classification is to arrange the data according to some common characteristics such as age, weight, place etc. The last step in organization is tabulation. Under tabulation the classified data are arranged in rows and columns so that there is absolute clearity in the data presented.

Presentation : Classification and tabulation make the data easy to understand and well arranged, after that they are ready for presentation. Data presented in an orderly manner facilitate statistical analysis whereas presentation in the form of diagrams and graphs so as to leave a permanent impression on the mind at a glance.

Analysis : After collection, organization and presentation the next step is that of analysis. Collected data are analysed by different statistical methods. Analysis is also done with relevance of object of investigation. Analysis should be such that can lead to conclusions easily. In analysis most commonly used methods are measures of central tendency, measures of variation, correlation, regression etc.

Interpretation : The last stage in statistical investigation is interpretation. Through analysis we get results for interpretation. Investigator should drawn correct inferences from data, otherwise the object will not be achieved. The interpretation of data is a difficult task and necessitates high degree of skill and experience. If the analysed data are not properly interpreted, the whole object of the investigation may be defeated and conclusion drawn also be fallacious. Correct interpretation will lead to a valid conclusion and one can take suitable decision.

Activity B

Comment on the following from the point of view of a statistical investigator :

(i) Data about cost of production in a company were collected.

_____

(ii) The data about amount of sales and profit of last two years were collected. The management wants to know about return on capital employed.

_____

(iii) Comparison about profitability of two companies was made. One company collected data of profit before tax and other collected data of profit after tax.

_____

(iv) Comparison of production of rice in two countries was made. One country was measured data in ton and other was in bushel.

_____

2.7 Executing the Survey

After planning of survey is completed the next step is to execute the survey. The various phases of the work subsequent to the planning stage may be enumerated as follows :

- Setting up an administrative organisation
- Design of forms

• Selection, training and supervision of the field investigators

• Control over the quality of the field work and field edit

• Follow-up of non-response

• Processing of Data

• Preparation of Report

Setting up an administrative organisation : The administrative organisation required for an enquiry will depend very much on the nature and scope of an enquiry. When the area of enquiry is wide then supervision through a central office is to be difficult and in that cases establishment of regional offices is best. If some organisations already exist, they also can be used for this purpose.

Design of form : Designing of various forms that will be used in the course of the enquiry should be given careful attention.

Selection, training and supervision of field investigators : Success of survey depends upon the field investigators. Thus, it is necessary that they are properly selected, trained and their work closely supervised. The important task of investigation is collection of data and it is done by field investigators or enumerators. The nature of the enumerator's job is such that great care has to be taken in his selection.

The enumerators should be honest, intelligent, and hard working. The enumerators should also be able to create friendly atmosphere to put the respondent. After selecting the enumerators they should be properly trained. The enumerators should know the purpose of survey . They should be acquainted with definitions of the terms used about the manner in which data are to be collected and how can interview be conducted. They should know the definitions of the terms used. The training may also be imported with the help of instruction manuals. The work of the enumerators is also to be watched carefully. The presence of supervisors in the field has a wholesome effect on enumerator's performance.

Control over the quality of the fieldwork and the field edit : This step must be taken to ensure that the survey is under statistical control. A system of field checks should be introduced. These checks done by supervisors and it should be carried out on a random sample. Investigators should not have any prior knowledge about checking. If it is found that the enumerator is not following the instructions, he should be removed from the field.

Before the questionnaires and schedules are passed on to the headquarters, they should be checked by supervisor. This editing is highly useful because there may be some wrong entries, omissions, inconsistencies and other errors.

Follow up on non-response : In spite of all best efforts, some of the respondents may not co-operate. A suitable method for dealing with those from whom the required information could not be obtained should be set up. After that supervisory staff can make vigorous efforts for obtaining response.

It should also be taken in the sight that enumerators are not allowed to make substitutions. If this practice is followed the enumerators will not take pains to persuade the non-respondent to co-operate and will introduce bias in the survey results.

Processing of data : After the data have been collected, the office work starts. The data are to be checked, coded and tabulated. These activities are as important as collection of data. There are chances of errors at every stage therefore one should be cautious. In editing process investigator should note that questionnaires are completed and information supplied is consistent and accurate.

The response obtain from edited questionnaires should be coded. To facilitate the analysis, responses are translated in numerical terms. For this purpose, list of code should be set up. After the completion of editing and coding analysis should by hand or by machine. Now-a-days most of survey work is tabulated by computers. By the use of computers investigators cannot save only his time but he also analyse the operation of complete system, which cannot be studied by other mean so economically.

Preparation of Report : After the analysis of data it is essential that the result of the survey should be presented in the form of a report. It is the final step in execution to prepare the report. Reports which are presented may be of two types **(i)** General Report**; (ii)** Technical Report

Those who are primarily interested in the results, general report giving a description of the survey for them whereas technical report giving details of sample design, procedure used for computation and accuracy etc.

The United Nations statistical office made some recommendations on the subject of the preparation of reports. According to these recommendations following point of the survey should be presented in general report :

• Statement of the purpose of survey.
• Description of the coverage
• Collection of information.
• Numerical results.
• Accuracy attained.
• Miscellaneous considerations: Such as period to which the data refer, time taken for field work, survey is an isolated or not, cost of the survey on different stages etc.

A technical report should highlight the following aspects:-

• Specification of the frame.
• Design of the survey
• Personnel and equipment
• Statistical analysis and computational procedure.
• Comparisons with other sources of information
• Observations of technicians.

It should be observed that utmost care needs by a survey at every stage. However, everything else is done well but poor work in any stage ruin a survey.

Activity C

Describe the procedure that you would adopt in conducting a survey of student's reading habit in a college.

_____

_____

_____

2.8 Advantages and Disadvantages of Investigation

Advantages:

• Investigation is an efficient way of collecting information from a large number of respondents. Statistical techniques can be used to determine validity, reliability and statistical significance.

• Surveys are flexible so a wide range of information can be collected. They can be used to study attitudes, values, beliefs and past behaviours.

• Surveys are standardized so they are relatively free from several types of errors.

• In surveys, the focus is provided on standardized questions in such a way that it may attain an economy in data collection. Only questions of interest to the researcher are asked, recorded, codified and analyzed, therefore time and money is not wasted.

Disadvantages

• Surveys depend on subjects, motivations, honesty, memory, and ability to respond. Subjects may not be aware of their reasons for any given action.

• The individuals chosen to participate in surveys are often randomly sampled, errors due to nonresponse

may exist. That is, people who have been chosen to respond on the survey may be different from those who do not respond, thus biasing the estimates.

• Survey question answer-choices may lead to vague data sets because at times they are relative only to a personal abstract.

2.9 Summary

The search for knowledge done by analysing numerical fact is called as a statistical investigation. Before starting any investigation a preparation of a systematic programme for the same is necessary otherwise, irrelevant data will give incorrect results and money, time and labour will also be wasted.

Investigation may be of different types such as direct and indirect investigation, official, semi-official and non-official investigation, regular and adhoc investigation, confidential and non-confidential investigation and initial and repetitive investigation.

Planning is essential before the investigations actually commence. The object is going to fail utterly in case the investigation has not been planned properly. **Nater and Waserman** states that the systematic order and accuracy of the data obtained in a survey depends directly on the carefulness observed during the planning of survey.

Data should be collected according to the plan and result can be obtained by using various methods of statistics. After completion of whole exercise a report should be presented. A report is a presentation of whole procedure of an investigation in short and concise form giving the essential information in brief descriptive of mode.

2.10 Key Words

Primary Data- Data collected by the investigator afresh for the first time to be used in an investigation.

Secondary Data - Data which were collected earlier by someone else and which are now in published or unpublished form

Coefficient - Rate per unit

Statistical Unit - An attribute or group of attributes conventionally chosen so that individual persons or items possessing those attributes may be enumerated.

2.11 Self Assessment Questions

Q. 1 What is statistical investigation? Explain in detail the necessary requisites that should be borne in mind while planning statistical investigation.

Q. 2 What do you mean by statistical unit? what are the various types of statistical units?

Q. 3 What do you mean by 'Reasonable Degree of Accuracy'? Explain with example.

Q. 4 Explain the steps of execution of survey.

Q. 5 Discuss the different stages of statistical investigation.

Q. 6 Write short note on:
(a) Degree of Accuracy
(b) Units of measurement
(c) Coefficient

2.12 Reference Books

1. Sancheti, D.C.; Kapur, V. K., Statistics (Theory, Methods & Applications).
2. Gupta, S.P. Statistical Methods.
3. Elhance, D.N. Elements of Statistics.

Unit - 3 Collection of Data

Structure of Unit:

3.0 Objectives

After reading this unit you will be able to understand:

- About data and of data types
- About methods of collecting primary data alongwith their merits and demerits
- About questionnaire and schedule
- Suitable method of collecting primary data
- Main sources of secondary data

3.1 Introduction

Data are foundation of any statistical investigation and data collection is the same for a statistician as collecting stones, mortar, bricks and household materials for a house builder. We can not build a house without collecting basic material, same as we cannot go further without collecting data. Utmost care must be exercised while collecting data because data constitute the foundation on which the superstructure of statistical analysis is built.

3.2 Collection of Data: Meaning and Definition

Data source is the main step in the process of statistical investigation. Data collection means searching of the desired facts and information. After planning statistical investigation the most important work for the statistician is to collect data. If data is available in books, magazines and newspapers, these data should be collected at a place or if they are not available then collecting them from the original field, is called collection of data.

"Collecting data relates to a purposeful gathering of information related to the subject matter of the study from the modules under investigates."

Collection of data is the basic process of scientific statistics because data are the main basics of investigation and its success totally depends upon data. Data should be accurate and sufficient so that correct conclusion

may be extracted. Whatever may be the method of collecting data, the results obtained from the analysis or properly interpreted and analysed. Hence, if the data are inaccurate and inadequate the results of analysis would be faulty and the decisions taken are misleading.

3.3 Types of Data

There are two types of data-

(1) Primary data
(2) Secondary data

(1) **Primary data :** Primary data are those data which have been collected by investigator from the original field and by a study specifically designed to fulfil the data needs of the problems. Primary data are freshly collected for a specific objective from starting to end. In other words, the data collected originally for the first time by an agency or an investigator for statistical investigation for example.

• The data collected by Ministry of Industries and made available through various publications constitute primary source.

• Data obtained in a population census by the office of the registrar general and census commissioner, Ministry of Home Affairs are primary data.

• Reserve Bank collects information about Home loan for the first time during an investigation, the data so collected will be primary data.

"Collection means the assembling, for the purpose of particular investigation of entirely new data, presumably not already available in published sources". **– Crum, Patton and Tebbutt**

(2) **Secondary data** : Data which are not collected originally but rather obtained from published or other unpublished sources are known as secondary data. Secondary data are already collected and published by various persons or agencies and then used by the investigator so we can say that this kind of data are not fresh and original. Secondary data are not collected by investigator by itself but collected by others and used by him. Investigator just use these data for his statistical investigation. The secondary data constitute the main material on which the statistical work is executed by investigators in many investigation.

One should go through secondary data before collecting primary data so that one can know that what has been already known and what not.

For example, in the case of examples stated in the primary data section. If the data collected in those examples are used by other investigators in their statistical investigation, then this type of data used is perfect example of secondary data.

Data is primary for the individual agency or institutions collecting them whereas the rest of the World they are Secondary.

3.4 Difference Between Primary and Secondary Data

The following are the main differences between primary and secondary data-

1. Primary data are original and they work as raw material in statistical methods where as secondary data are not original, they are like finished material.

2. Primary data are collected according to the objective of the investigation but in secondary data necessary changes have to be made to make them useful as per the requirement of the objective of the investigation.

3. More hard work, time and money is needed in the primary data collection, whereas less hard work time and money is required in secondary data collection.

4. Primary data are collected by the investigator himself or his agents but secondary data were collected and published earlier by some other persons or institutions.

5. Data which are primary in the hands of one become secondary in the hands of another.

Activity A:

State whether the following data are primary or secondary.
(i) The chief executive is preparing a report on the prospect of water requirement in the district head quarters city for next 5 years from the data available from the "water management" publication by his government.

_____

(ii) A group of college students is examining the relationship between Tobacco Chewing and mouth cancer on the basis of data published in "The Journal of Medicine".

_____

(iii) JET Airlines writing a report on accidents caused by birds stricks during 2008-2009 using the data available in "The Annual Report of the Chief Safety Officer of The JET Airlines" Published by the Airlines.

_____

(iv) Reserve Bank Collects information about Home Loan for the first time during an investigation.

_____

(v) The data collected by Ministry of Agriculture and made available through various publication.

_____

3.5 Methods of Collecting Primary Data

Primary data may be obtained by applying any of the following methods:

- Direct oral investigation;
- Indirect oral investigation;
- Information through local sources or correspondents;
- Mailed questionnaire;
- Schedules sent through enumerators.

Direct Oral Investigation : under direct oral investigation method the investigator goes directly to the investigation field and directly contacts with the informants. Direct contact means face to face contact with the person from whom the information to be obtain. According to Prof. W.A. Neiswanger, "It is better to observe and record behaviour than to ask a question".

In this particular method same special qualities must be in the investigator like he/she should be practical, expert, polite, impartial, fine observer and patient.

Suitability:

This method is suitable for intensive rather than extensive investigation. Hence, it should be used only in those cases where intensive study of limited field is desired-

1. When originality and accuracy of data is required.

2. When investigation field is limited

3. When data needed to be kept secure

4. When qualities like expert, polite, practical, patient, impartial and fine observer are needed in investigator.

For example family Income & Expenditure, living standard of workers, education and unemployment in youth in a particular (or Limited) area.

Merits:

(i) **Originality and accuracy :** Due to the presence of investigator in the field, there is high rate of originality and accuracy. Personal presence and personal contact makes the data more original and more accurate.

(ii) **Reliable information :** The investigator extracts much more information also other than those about the main topic and these are completely reliable.

(iii) **Flexibility :** This method is flexible. The investigator, according to the requirement can get desired information by just changing questions he can twist the questions keeping in mind the informant's reaction.

(iv) **Uniformity :** There will be uniformity in collected material and there will be no chance of ambiguity because the same person has collected the data. Due to uniformity data are comparable.

(v) **Correctness :** In this method, personal presence of investigator makes the data more reliable and correct.

(vi) **Suitability :** This method is more suitable for intensive investigation.

Demerits

(i) **Limited area :** Direct oral investigation method can only be applied in limited area. It is not suitable for wide area.

(ii) **Partiality :** Though it is expected of the investigator to be perfectly impartial yet human nature gets effected by individual feeling like partiality.

(iii) **Wastage :** This method is more expensive, time consuming and hard work demanding.

(iv) **Subjective result :** The major draw back of this method is that it is totally subjective in nature. The success of the investigation indirectly depends upon skill, tact, diplomacy, intelligence of the investigator. If investigator does not have these qualities the results are not reliable.

Indirect Oral Investigation : In Indirect oral investigation method information is not directly obtained from persons who are directly associated with the investigation rather than it is obtained with the persons, indirectly related with investigation as third party, known as witness. Witness are the persons who are close related to persons, concerned with the investigation who can provide authentic information. Witness are asked questions orally and their answers are recorded. A small list of questions pertaining to the subject matter of the inquiry is prepared.

For example: Information related to the workers are not collected from workers inspite of it is asked from worker's association, mill owners.

Suitability

1. Where investigation area is wide.
2. When direct Contact is not possible with informants.
3. When it is not considered right to ask questions from persons related to the investigation or to keep it secret.
4. This method is generally used by Government for setting up commission and committees for inquiry.
5. When data are complex in nature.

Merits

(i) **Wide area :** Wide are can be covered in indirect oral investigation method.

(ii) **Simplicity :** This method is more easy and convenient because work is done more rapidly without any problem and difficulty.

(iii) **Economic :** This method saves, time, money and energy.

(iv) **Impartiality :** Indirect oral investigation method is less influenced by partiality feeling of humans.

(v) **Secrete Information :** Under this method secrete information can also be obtained, which the informants do not want to disclose by asking third persons.

(vi) **Opinion of expert :** expert's opinion related to the investigation is obtained in this method. Opinion and suggestion of specialist are obtained incidentally.

Demerits:

(i) **Lack of accuracy :** This particular method provides inaccurate data comparatively to the direct oral investigation.

(ii) **Partiality in views :** It is possible that the witness can give wrong information because of carelessness, partiality or ignorance. This effects the result of the investigation.

(iii) **Lack of uniformity :** uniformity is lost because data are obtained by various informants of different nature. Therefore, data are not comparable.

Information through local sources or correspondents

In this method the investigator appoints local agents or correspondents in different places to collect information. These correspondents collect and transmit the information to the central office, where it is processed. These correspondents send information time to time on the basis of experience, choice, decision or estimation and observation.

"This method is useful when figures are required cheaply and expeditiously and accuracy is not prime importance." **– L.R. Konnor**

The special advantage of this method is that it is cheap and appropriate for investigation. However, it may not always insure accurate results because of the personal prejudice and bias of the correspondents.

Suitability

1. This method is generally adopted in these cases where the information is to be obtained at regular intervals from wide area.
2. This method is basically used by new papers, magazine and Radio stations etc.
3. This method is more suitable in these investigations which need less accuracy.

Merits

(i) **Wide area :** This method is more suitable for the wide area for collecting information.

(ii) **Economic :** Method saves time, money and energy.

(iii) **Estimated enumeration :** The method is more useful when estimated enumeration is needed expeditiously.

Demerits

(i) **Lack of accuracy and originality :** Because it is generally sent on the basis of estimate or experience therefore, the collected material lacks originality and high degree of accuracy.

(ii) **Lack of Uniformity :** There are different ways of correspondents and different basis of collecting information, as a result data are not uniform.

(iii) **Delay :** sometimes correspondents send information not at regular intervals, therefore importance of information is decreases.

(iv) **Biasness :** Collected material has direct effect of personal prejudice or partiality of the correspondent. This makes the data inaccurate and false.

Mailed Questionnaire Method: Mailed questionnaire method is a method in which a list of question related to the investigation is prepared and sent to the various informants by post called questionnaire.

Informants fill these questionnaire and return before due date Questionnaire includes covering letter and requesting letter having objectives of investigation and assurance to keep the information secrete. At the time of preparing questionnaire same points should be in consideration like questions are easy, small, clear and less in number having no confusing words or other frustative words. The question should be directly related with investigation and answers should be in Yes or No.

Suitability

1. Method is better suitable for wide area where informants are educated
2. Particularly opinion survey and consumer aptitude surveys use this kind of method.
3. This method is generally perfect for personal interview or face to face interview.

Merits

(i) **Economic :** Information of investigation related to the wide areas are easily, cheaply and timely available.

(ii) **Wide area :** Mailed questionnaire method is extremely useful for the wide area investigation.

(iii) **Originality :** The information regarding to the investigation are directly collected from the informants, so quality of originality became its main feature.

(iv) **Time Saver :** All questionnaires are mailed simultaneously, the replies are received soon therefore a lot of time can be saved.

(v) **Reduced Errors :** Information is given by informants themselves, so there are least chances of errors or inaccuracy.

Demerits

(i) **Incomplete and ambiguous :** Many a times informants do not return the questionnaire or if send the questionnaire is incomplete.

(ii) **Limited Scope :** This type of methods scope is only upto educated or literate persons.

(iii) **Lack of accuracy :** High degree of accuracy can not be expected in this method because questionnaire are not prepared carefully or questions are taken in wrong meaning or if partiality exists in informants then data became inaccurate.

(iv) **Not Practical :** In some cases informants hide their personal information like income, assets etc.

(v) **Lack of flexibility :** Method is not flexible because it is not possible to make necessary changes in questions when information received is incomplete.

Schedules sent through enumerators:

There are many limitations and problems in mailed questions method because of incomplete information, inaccuracy, half or partially filled questionnaire, so for removing all these problem of mailed questionnaire method, schedules sent through enumeration method is used. In this method questionnaire is filled in by the enumerators themselves by visiting the investigation places personally and asking questions.

In this way this method and last method has some similarities and basic dissimilarities, in last method questionnaire are filled by informants where as in this method information filled in schedules by enumerator. The list of question is known as schedule rather than questionnaire. The investigator appoints enumerator to go different places and collect information in schedules.

Suitability:

1. Method is used in large business organizations, public corporations, research organizations, Government and private organizations, In India census population investigation is done by this method.
2. Where the area or scope of investigation is very wide and time, money and labour are also available.

Merits:

(i) **Wide area :** This method is suitable for wide area and for both educated and uneducated.

(ii) **Accuracy :** Experienced, trained and eligible for research enumerators are appointed for filling in schedule.

(iii) **Reliability :** Due to direct contact answers of typical questions can be obtained.

(iv) **Impartiality :** There is no influence of partiality because there is both type of enumeration, some are in favour and some are against of it.

Demerits:

(i) **Experience :** Method requires more experience on the appointment and training of the enumerators

(ii) **Training problem :** It needs much time and money to train enumerators appointed for the job.

(iii) **Supervision :** To check the working of enumerator is not an easy job. In absence of supervision data collected may be inaccurate.

3.6 Drafting the Questionnaire

Questionnaire is a set of questions pertaining to survey which is sent to the respondents for filling in by them in their own handwriting.

"Questionnaire refers to a device for securing answers to questions by using a form which the responded fills in himself". – **Goode and Hatt**

"In drafting of a suitable questionnaire special care and caution must be taken so that all relevant and essential information regarding survey may be collected without any difficulty, ambiguity and vagueness. The success of the survey depends on the quality of the questionnaire" – **Arthor Kornhauser**

The following general points are essential in a good questionnaire :-

1. **Covering Letter** : A covering letter from the organization of the enquiry should be enclosed with the questionnaire with the following reasons : (a) Person conducting the survey must introduce himself and state the objective of the survey (b) It should contain a note regarding the operational definitions to the various terms and concepts used in the questionnaire (c) ensure the informants to keep secrete the information extracted by them (d) If respondent wants a copy of the results of the survey, provide him (e) Quick and better response provide him awards, gifts and incentives (f) It should take the respondent in confidence.

2. **Size of questions should be small :** The Number of questions should be kept to the minimum, keeping in view the nature scope and objective of the inquiry. 15 to 25 may be regarded as a fair number.

3. **Questions should be easy and clear to understand :** Questions should be clear, ambiguous, non-offending simple so that the informant can easily answer the question.

4. **Personal & Confidential question should be avoided :** Personal and confidential questions like Income and Assets should be not asked because informants try to hide these kind of details. Even if compulsory then ask such questions at the end of the questionnaire when the informant feel more at ease.

5. **The questions should be in a logical manner :** Questions in a questionnaire should be arranged logically they should not skip back and forth from one topic to another e.g. it would be illogical to ask a man his income before asking him whether he is doing a job or not. Thus the questions should be in sequence questions about identification and description should come first and after that questions about major information should come. Opinion based questions should be placed at the end of the questions.

6. **Ambiguous questions ought to be avoided :** Ambiguous question means different thing to different people. The use of unpopular and complex words should be avoided. For example instead of asking a student "Are you a good sportsmen?" They are asked "have you participated in sports ever?"

7. **Type of Question**

Multiple choice or objective question : Questionnaire must contain questions like multiple choice or objective question because they are simple to answer and easy to understand.

Yes/No Questions: As far as possible the questions should be of such a nature that they can be answered easily in Yes or No. These answers are quick.

One word questions : These type questions are direct and easy to answer. Informants does not hesitate in answering one word questions e.g. what is your annual income what is your name?

Fill in the blank : Questions may be in the form of fill in the blank, informant just have to fill the information in the blank space provided e.g. subscribed to daily news paper-

Short Answer questions : There are many question in questionnaire whose answers are not possible in one word but they should be answered as brief as possible e.g. why do you use a particular brand of bath soap as compared to other brands?

Open question : Open questions are those questions in which no alternative answers are given and the informants are free to express answers according to their feelings and independent opinions on the problem in their own words e.g. what are the drawback in our examination system?

8. **Attractive questionnaire :** This questionnaire should be attractive in the form of good paper, its layout and the format of the questionnaire, multi colouring etc.

9. **Cross-Checks** : The questionnaire should be prepared to provide internal checks on the accuracy of the information given by the respondents by asking some connected question with respect to fundamental to the enquiry.

10. **Method of Tabulation :** The method to be used for tabulating the result should be determined before the final draft of questionnaire is made. It may be hand written computerized or machine operated.

11. **Instruction to the informants :** The questionnaire should specify the time within which it should be sent back and the address at which it is to be post. Instruction about unit of measurement should also be given.

Activity B:
You are the Sales Promotion officer of Alpha Cosmetics Co. Ltd. Your company is about to market a new product. Design a suitable questionnaire to conduct a consumer survey before the product is launched.

_____

_____

_____

3.7 Schedule

Schedules is also a list of question, used for the collection of primary data. In schedule questions are asked and filled in by the enumerators, in a face-to-face situation with another person (respondents).

3.8 Difference between Questionnaire and Schedule

Following are the differences between schedule and questionnaire :-

1. Questionnaire is to be filled in by the respondents themselves whereas schedule is filled in by the trained enumerators.

2. Questionnaire can be used only where the informants are literate but schedule can be used everywhere whether informants are literate or illiterate.

3. Questionnaire method is economical because information are obtained by post. Schedule method is comparatively expensive because enough money is required for training, pay and allowances of enumerators.

4. Questionnaire method usually takes more time to obtain the information whereas in schedule method information is collected by enumerators personally, therefore no question of delay.

5. Basis of success of the questionnaire method depends upon the selection of questions and co-operations of the respondents. In schedule method the success depends upon the training and performance of the enumerators.

6. The standard of accuracy and reliability of the information received is not very high in questionnaire method but in schedule method accuracy and reliability of the information is comparatively much high.

7. Since questionnaire is to be sent through mail so there is no direct personal contact with the respondents but schedules are to be filled in by the enumerators so there is a direct personal contact can be established,

8. No guideline is sent with a questionnaire while a list of guidelines is attached with the schedule to be filled in by the enumerations.

3.9 Choice of a Suitable Method of Collecting Primary Data

After reviewing various methods of collecting primary data, the question arises which is the appropriate method that can be adopted in all circumstances? According to Dr. A.L. Bowley, "In collection and tabulation common senses is the chief requisite and experience the chief teacher." Therefore, it is very difficult to select the best method. Selection of a suitable method depends upon the following points:-

1. **Nature of investigation –** What method should be adopted for collecting primary data depends upon nature of investigation. If the area of investigation is limited and topic of investigation is serious, direct oral investigation is must suitated. If not possible then indirect oral investigation should be adopted. If informant are uneducated and area is wide then information can be collected through schedules filled in by enumeratiors.

2. **Object and Scope –** In limited areas the use of direct personal investigation method is possible but in a wide area questionnaire will have to be sent normally. For intensive investigation with a wide scope information can be obtained through schedules filled in by enumerator.

3. **Financial sources –** If financial recourses estimated, before starting investigation, then the suitable method according to the financial recourses can be set. Direct investigation and investigation through enumerator cost much where as question is less expensive.

4. **Desired accuracy** - Method of collection should be selected according to the degree of accuracy of the data is desired. Direct investigation in limited area have highest accuracy. Indirect investigation in wide area may not have much accuracy.

5. **Time -** If the investigator has to collect information within a short time, it is possible only through correspondents or questionnaire filled in by informants. If time available is adequate direct personal investigation can be adopted.

Activity C:

Which is the most suitable method of collection of data for the following purposes and why?

1.Information is to be obtained regarding the health to infants of a particular period in Jaipur.

_____

2. The study of income and expenditure of 500 workers in a plastic factory.

_____

3. The financial Budget of Indian Government of 2008-2009 is to be analysed by a researcher.

_____

4. A study is to be made of the changes in wholesale prices of sugar in India in different years taking for the year 2005 as the base year.

_____

5. The Chief Executive of SBBJ desires to find out the reasons for non payment of loans taken by a number of people from rural area.

_____

3.10 Sources of Secondary Data

Secondary data may be obtained from published or unpublished various sources:

I. **Published Sources-** Government and non government organizations, other investigators collect primary data on different subjects. This primary data has been used as a secondary data by other investigator.

(i) International Publications – International bodies and foreign Government's publications are used as secondary data. eg. U.N. Statistical Year Book, Demo graphic year book, Annual Reports of the IMF, I.L.O., W.H.O. and ASEAN SAARC etc.

(ii) Government Publication - Different departments established under various central and state Govt. ministries Govt. directorate and statistical divisions collect and published statistics, regularly. This type of data are more reliable. eg. Five year plan progress reports, census reports, RBI Bulletin, Economic Survey, Statistical abstract of India (Annual).

(iii) Semi Govt. Publications – Semi Government Publications like municipalities, district board, Panchayat etc. Publish figures regularly, like Birth & death data, education, health Reports publishers regularly.

(iv) Publications of non – Government institutions – Business institutes like FICCI, Hindustan Liver Ltd., NCAER, India statistical institute.

(v) Magazines & News papers – Economic Times, Financial Express, Commerce, Eastern Economist and The Chartered Accountant.

(vi) Publication of Individual Investigators – Many investigators Publish their research or primary data for public use.

(vii) Research organization – Indian statistical institute (ISI) Economic development research organization and national sample survey organization.

II. **Unpublished Sources:-** Some of the important data not yet published or left over in office files, documents or registers etc. This kind of unpublished statistical, may be useful for several investigation.

3.11 Precaution in the Use of Secondary Data

Secondary data should be used with extra caution since secondary data have already been obtained, it is highly desirable that a proper analysis, scrutiny or filtration of such data is made before they are used by investigator.

In the words of **L.R. Connor** "Statistics, especially other people's statistics, are all of pitfalls for user."

"It is never safe to take the published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticize the arguments that can be based upon them."

– Prof. A.L. Bowley

In using secondary data we should take a special note on the following points:-

- Whether data is suitable for investigation or not.
- Whether data is enough to fulfill the demands of investigation.
- Data must be reliable because unreliable data may generate false and bias results.
- Secondary data should be selected in accordance with the objective and basic requirement of the investigation.

Quite often secondary data do not satisfy immediate needs because they have been collected for other purpose. Even when directly pertinent to the subject understudy, secondary data may be just enough off the point to make then of little or no use.

Hence, in order to arrive at conclusions free from limitations and inaccuracies, the published data i.e. the secondary data must be subjected to thorough scrutiny and editing before it is accepted for use.

3.12 Summary

Collection of data is the first step of original process in statistical investigation collection of data means to make available necessary material in accordance with the objective of investigation. Basically data are of two type primary data and secondary data. Data collected from original field freshly for the first time is called primary data, if already collected primary data are used in investigation then it is called secondary data.

For the collection of primary data in relation to the situation we mainly use five methods Direct oral investigation, Indirect oral investigation, Information through local sources or correspondents, Mailed questionnaire, Schedule sent through enumerators, secondary data are collected from published sources International Publication, Government Publication, Semi Govt. Publications, Publications of non Government Institution, Magazines & News papers, Publication of Individual Investigator, Research Organization) and unpublished sources.

3.13 Key Words

Primary Source : It is one that itself collects the data.

Secondary Source : It is one that makes available data collected by some other agency.

Enumerator : Enumerator or investigator is a person who collects the information.

Respondent : A person who fills the questionnaire or supplies the required information.

3.14 Self Assessment Questions

1. What are the different methods of collecting statistical data? Which of these is most reliable and why.
2. Differentiate the following:

   (i) Primary and Secondary data
   (ii) Questionnaire and Schedule.
3. What do you understand by 'Secondary Data'? Mention the various sources of secondary data.
4. What is a questionnaire? What are the chief requirements of a good questionnaire for use in statistical inquiry?
5. What precautions are to be taken before collecting secondary data?

3.15 Reference Books

1. Gupta, S.P. Statistical methods.
2. Kothari, Research methodology.
3. Yadav, Jain, Mittal, Business statistics.

Unit - 4 Classification and Tabulation of Data

Structure of Unit:

4.0 Objectives

After completing this unit, you will be able to :

- Define Classification & Tabulation.
- Assess the Objects and Characteristics of Classification
- Explain various methods of Classification.
- Evaluate the difference between Classification and Tabulation.
- Discuss various types of Statistical Series.
- Excess various parts of Tabulation & Rules for Tabulation.

4.1 Introduction of Classification

After the collection and editing of data the next stage is of classification. Classification means arranging the data in different classes according to their qualities and characteristics. So, the process of arranging data in groups or classes according to resemblances and similarities in technically called classification. Classification is the process of grouping data into homogeneous classes and categories. It is necessary to provide the collected data a form or structure to be appropriate for analysis and draw inferences. To serve the purpose, the collected data have to be classified into classes and sub-classes according to their characteristics. This process is called 'Classification'. According to Prof. L.R. Connor, "Classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblances and affinities and gives expression to the unity attributes that they may subsist among a diversity of individuals."

Similarly A.M. Tuttle has said, "A classification is a scheme for breaking a category into a set of parts, called classes according to some, precisely defined differing characteristics possessed by all the elements of category." According to Gregory and Ward, "Classification is the process of relating the separate items within the mass of data we have collected."

4.2 Objects of Classification

The main objectives of classification are as under :

1. **To Clarify Similarity and Dissimilarity** : The mass of figures collected is easily arranged into few classes having common characteristics. For example, facts collected and having similar characteristics are placed in one class such as students passed and failed, literate and illiterate persons, married and unmarried persons etc.

2. **To Facilitate Comparative Study** : Classification enables us to make meaningful comparisons between variables. For example, classification of students according to their divisions, (I, II, III etc.) of the two colleges will make comparison possible about the intelligence of the students of those colleges.

3. **To Make the Data Easy and Concise** : Classification presents the huge unwieldy raw data in a condensed form so that it may be easily understandable and highlights the main features contained in the collected data. For example, the marks obtained by 2000 students of a college given separately are useless to infer any conclusion. But if it is classified into the groups of first, second and third division is and the failures, is easily comprehensible to the mind from which conclusions can be drawn.

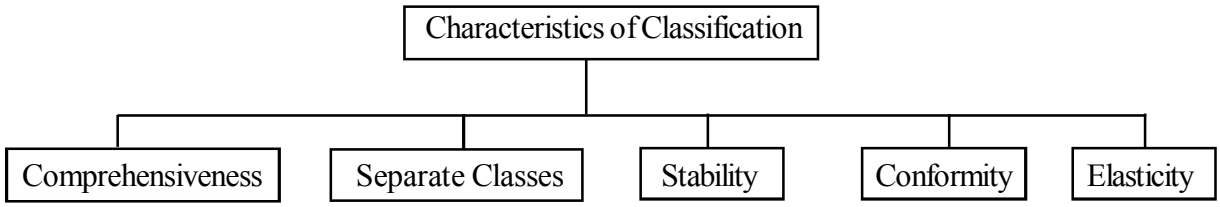
4. **To Make Data Scientific and Understandable** : Classification enables to present data in an organised and scientific manner so as to draw desired conclusions. For example, the students of a college if classified according to faculty (Science, Commerce, Arts and Law), sex (male and female) and class and presented accordingly are easily understandable.

5. **To Make Data Concise** : Statistical data collected during the course of enquiry are so varied that it is not possible to infer any conclusions even after a careful study. Classified data can be easily understood for the purpose of analysis and interpretation.

6. **To Provide Base for Tabulation** : Only classified data can be presented in a tabular form. Thus, it provides the basis for tabulation.

4.3 Characteristics of Classification

A good classification should have various characteristics. These are shown as follows -



4.4 Methods of Classification

The methods or the criteria w.r.t. which the data are classified primarily depend on the objectives and the purpose of the enquiry. Generally, the data can be classified on the following four methods:

(i) Geographical i.e., Area-wise or Regional.
(ii) Chronological i.e. w.r.t. occurrence of time.
(iii) Qualitative i.e., w.r.t. some character or attribute.
(iv) Quantitative i.e. w.r.t. numerical values or magnitudes.

In the following section we shall briefly discuss them one by one:

(i) **Geographical Classification** :- In this classification the basis of classification is the geographical or locational differences between the various items in the data like States, Cities, Regions, Zones, Areas etc. For example, the yield of agricultural output per hectare for different countries in some given period or the density of the population (per square km.) in different cities of India, is given in the following tables.

Table-4.1

Agricultural Output of Different Countries
[In kg. Per Hectare]

Country	Average Output
India	120
USA	580
Pakistan	260
USSR	730
China	270
Syria	615
Sudan	338
UAR	750

Table-4.2

Total of Population
In Different Cities of India

City	Total of Population (in Lakh)
Kolkata	40
Bombay	50
Delhi	45
Madras	40
Chandigarh	30

(ii) **Chronological Classification** : Chronological classification is one in which the data are classified on the basis of differences in time, e.g., the production of an industrial concern for different periods; the profits of a big business house over different years; the population of any country for different years. We give below the population of India for different decades.

Table-4.3

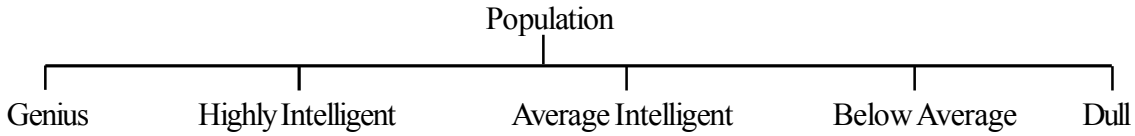
Population of India

Year	Population (in Crore)
2003	76
2004	80
2005	86
2006	90
2007	95
2008	100
2009	120
2010	135

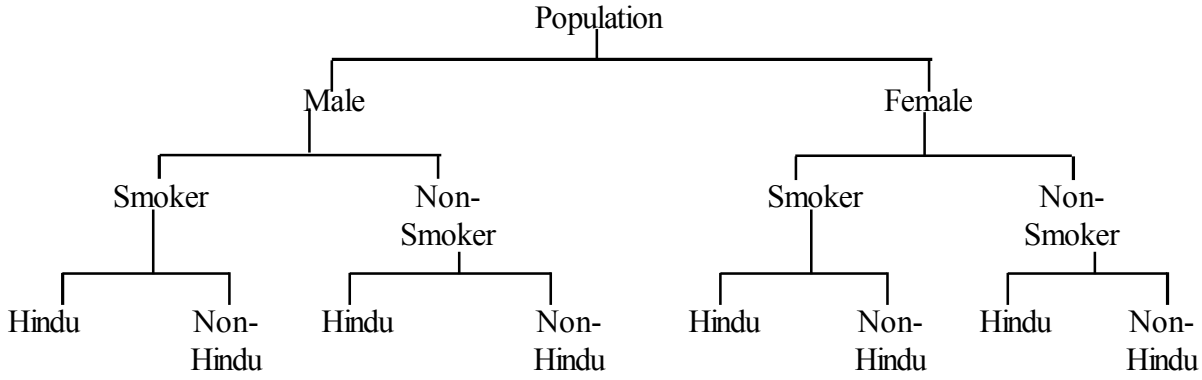
The time series data, which are quite frequent in Economic and Business Statistics are generally classified chronologically, usually starting with the first period of occurrence.

(iii) **Qualitative Classification** : When the data are classified according to some qualitative phenomena which are not capable of quantitative measurement like honesty, beauty, employment, intelligence, occupation, sex, literacy, etc., the classification is termed as qualitative or descriptive or w.r.t. attributes. In qualitative classification the data are classified according to the presence or absence of the attributes in the given

units. If the data are classified into only two classes w.r.t. an attribute like its presence or absence among the various units, the classification in termed as simple or dichotomous. Examples of such classification are classifying a given population of individuals as honest or dishonest; male or female; employed or un-employed; beautiful or not beautiful and so on. However, if the given population is classified into more than two classes w.r.t. a given attribute intelligence the various classes may be, say, genius, very intelligent, average intelligent, below average and dull as given below :



Moreover, if the given population is divided into classes on the basis of simultaneous study of more than one attribute at a time, the classification is again termed as manifold classification. As an illustration, suppose we classify the population by sex into two classes, males and females and each of these two classes is further divided into two classes w.r.t. another attribute say, smoking i.e., smokers and non-smokers, thus giving us four classes in all. Each of these four classes may further be divided w.r.t. a third attribute, say, religion into two classes Hindu, Non-Hindu and so on. the scheme is being explained below.



(iv)**Quantitative Classification** : If the data are classified on the basis of phenomenon which is capable of quantitative measurement like age, height, weight, prices, production, income, expenditure, sales, profits, etc., it is termed as quantitative classification. The quantitative phenomenon under study is known as variable and hence this classification is also sometimes calls classification by variables. For example, the earnings of different wholesale stores may be classified as under :

Table-4.4

Daily Earnings (In Rupees) of 100 Wholesale Stores

Daily earnings (Rs. in 000')	Number of stores
Up to 100	6
101-200	14
201-300	18
301-400	30
401-500	18
501-600	6
601-700	4
701-800	4
	100

4.5 Types of Statistical Series

There are three types of Statistical Series. These are as follows :-

(1) **Individual Series** : If every item or unit which we are studying, is put separately then the distribution is known as individual series. In this case every item is independent. No item is placed under any group and is kept fully independent.

Example-1

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks (Out of 50)	8	17	28	17	23	30	20	9	35	47

Example - 2

Financial Year	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10
Output ('000 tons)	10	15	18	16	17	20	25

In example (1) above ten students have been measured in terms of marks obtained. First student has secured 8 marks, while second 17, third 28 and so on. The marks of each student have been written against his serial number.

In example (2) output (in '000 tons(for different financial years are given. It is a time series. As the output is given for each unit of time individually, it is called individual series.

Arraying the Data : When in case of an individual series, ungrouped data are written either in increasing or decreasing order, it is called 'to array'. In ascending order, minimum value is written first, then next higher value and so on. The maximum value will appear at the last. While in descending order the reverse order is followed. The data of example (1) above will be arranged as under :

Ascending Order		**Decending Order**	
S.No.	Marks	S.No.	Marks
1	8	1	47
2	9	2	35
3	17	3	30
4	17	4	28
5	20	5	23
6	23	6	20
7	28	7	17
8	30	8	17
9	35	9	9
10	47	10	8

(2) **Discrete Series** : Discrete series is one in which exact measurement are possible in whole numbers. Every item has its own importance. In such a case size or value as well as items both are given. In the words of boddington, "a discrete series is one in which individual value differ from each other by definite amount."

The following are some examples :

Example-3

No. of Children (x)	1	2	3	4	5	6	7	8	
No. of Familes (f)	4	5	8	13	3	2	1	1	**N = 37**
(cf)	4	9	17	30	33	35	36	37	

Example-4

No. of words (size)	1	2	3	4	5	6	7	8	9	
No. of letters (freq.)	9	5	5	9	4	2	4	1	1	Total 40

(3) **Continuous Series** : In this series the size of items is not definite but it lies between the numbers. This becomes necessary in case of such variables which can take any fractional value and in whose case an exact measurement is not possible. According to Prof. Boddington, "In continuous series variables can take any value between maximum and minimum values." The following is an example of such series :

Example-5

Weights (in Kg.)	40-45	45-50	50-55	55-60	60-65	65-70
No. of Students (f)	17	10	23	11	19	60

4.6 Meaning and Definition of Tabulation

By tabulation we mean the systematic presentation of the information contained in the data, in rows and columns in accordance with some salient features or characteristics. Rows are horizontal arrangements and columns are vertical arrangements. In the words of A.M. Tuttle : "A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings, and notes to make clear the full meaning of data and their origin."

Professor Bowley in his manual of Statistics refers to tabulation as "the intermediate process between the accumulation of data in whatever form they are obtained, and the final reasoned account of the result shown by the statistics".

4.7 Importance of Tabulation

Tabulation is one of the most important and ingenious device of presenting the data in a condensed and readily comprehensible form and attempts to furnish the maximum information contained in the data in the minimum possible space, without sacrificing the quality and usefulness of the data. It is an intermediate process between the collection of the data on one hand and statistical analysis on the other hand. In fact, tabulation is the final stage in collection and compilation of the data and forms the gateway for further statistical analysis and interpretations. Tabulation makes the data comprehensible and facilitates comparisons (by classifying data into suitable groups) and the work of further statistical analysis, averaging, correlation etc. If makes the data suitable for further Diagrammatic and Graphic representation.

4.8 Objects of Tabulation

Some of the major objectives of tabulation are mentioned below :

(1) **To Make Data Concise** : The mass quantitative information is put in condensed and proper form with the help of a table. Thus the scattered data are put together for analysis and interpretation.

(2) **To Simplify Statistical Data** : The main object of tabulation is to condense the mass and complicated numerical information to make it easily understandable.

(3) **To Clarify similarity and Dissimilarity** : Tabulation clarifies the similarities and dissimilarities of data since all facts of some attributes are put together.

(4) **To facilitate Comparison** : Tabulation facilitates quick comparison of data shown in rows and columns.

(5) **To Present Facts in Minimum Space** : Tabulation presents the facts in minimum of space and conveys information in a far better manner than textual material.

(6) **To Clarify the Characteristics of Data** : When data are presented in the form of a table, there is no need of giving wordy explanation. It clearly explains the chief characteristics of data.

(7) **To Detect Errors in the Facts** : When the facts collected are presented in a table, some important omissions are detected and factual errors can be rectified.

4.9 Difference between Classification and Tabulation

There are a few differences between these two processes, which are as follows :

(1) **Priority or Order**: First collected data are classified and then the same are presented in tables. Thus classification forms the basis for tabulation. Both processes are to be performed in sequence.

(2) **Basis** : Under classification, data are classified into various classes according to their similarities and dissimilarities; whereas in tabulation, classified data are presented in columns and rows.

(3) **Object** : Classification is the process of analysis whereas tabulation is the process of presenting data in orderly and suitable manner.

(4) **Presentation** : Thus classification tends to classify the statistical data into classes and subclasses, while tabulation presents classified data under appropriate headings and sub-headings.

4.10 Parts of Tabulation

The various parts of tabulation vary from problem to problem depending upon the nature of the data and the purpose of the investigation. However, the following are a must in a good statistical table :

(i) Table number
(ii) Title
(iii) Head notes or Prefatory notes
(iv) Captions and Stubs
(v) Body of the tables
(vi) Foot note
(vii) Source note.

(i) **Table Number** :- If a book or an article or a report contains more than one table then all the tables should be numbered in a logical sequence for proper identification and easy and ready reference for future. The table number may be placed at the top of the table either in the centre above the title or in the side of the title.

(ii) **Title** :- Every table must be given a suitable title, which usually appears at the top of the table (below the table number or next to the table number). A title is meant to describe in brief explanatory. It should precisely describe the nature of the data (criteria of classification, if any); the place (i.e. the geographical or political region or area to which the data relate); the time (i.e. period to which the data relate) and the source of the data. The title should be brief but not an incomplete one and not at the cost of clarity. It should be un-ambiguous and properly worded and punctuated. Sometimes it becomes desirable to use long titles for the sake of clarity. In such a situation a catch title may be given above the 'main title'. Of all the parts of the table, title should be most prominently lettered.

(iii) **Head note (or Prefatory notes)** :- If need be, head note is given just below the title in a prominent type usually centered and enclosed in brackets for further description of the contents of the table. It is a sort of a supplement to the title and provides an explanation concerning the entire table or its major parts-like captions or stubs. For instance, the units of measurements are usually expressed as head such as 'in hectares', in millions', in quintals', in Rupees', etc.

(iv) **Captions and Stubs** :- Captions are the headings or designations for vertical columns and stubs are the headings or designation for the horizontal rows. They should be brief, concise and self explanatory. Captions are usually written in the middle of the columns in small letters to economise space. If the same unit is used for all the entries in the table then it may be given as a head note along with the title. However, if the items in different columns or rows are measured or expressed in different units, then the corresponding units should also be indicated in the columns or rows. Relative units like ratios, percentage etc., if any, should also be specified in the respective rows or columns. For instance, the columns may constitute the population (in millions) of different countries and rows may indicate the different periods (years).

Quite often two or more columns or rows corresponding to similar classifications (or with same headings) may be grouped together under a common heading to avoid repetitions and may be given what are called sub-captions or sub-stubs. It is also desirable to number each column and row reference and to facilitate comparisons.

(v) **Body of the Table** :- The arrangement of the data according to the descriptions given in the captions (columns) and stubs (rows) forms the body of the table. It contains the numerical information which is to be presented to the readers and forms the most important part of the table. Undesirable and irrelevant (to the enquiry) information should be avoided. To increase the usefulness of the table, totals must be given for each separate class/category immediately below the columns or against the rows. In addition, the grand totals for all the classes for rows/columns should also be given.

(vi) **Foot Note** :- When some characteristic or feature or item of the table has not been adequately explained and needs further elaboration or when some additional or extra information is required for its complete description, foot notes are used for this purpose. As the name suggests, foot notes, if any, are placed at the bottom of the table directly below the body of the table. Foot notes may be attached to the title, captions, stubs or any part of the body of the table. Foot notes are identified by the symbols *, **, ***, @ etc.

(vii) **Source Note** :- If the source of the table is not explicitly contained in the title, it must be given at the bottom of the table, below the foot note, if any. The source note is required if the secondary data are used. If the data are taken from a research journal or periodical, then the source note should give the name of the journal or periodical along with the data of publication, its volume number, table number (if any), page number etc, so that anybody who uses this data may satisfy himself, (if need be), about the accuracy of the figures given in the table by referring to the original source. Source note will also enable the user to decide about the reliability of the data since to the learned users of Statistics the reputations of the sources may vary greatly from one agency to another.

4.11 Rules for Tabulation

There cannot be any hard and fast rules for tabulation similar in all cases, yet certain rules and procedures may be laid down for guidance to prepare a table.

(1) **Number** : Each statistical table must have a number so that it may be identified easily.

(2) **Title** : A proper title of a table must be placed above it. A good title is one which is properly worded expressing clearly the nature of data contained. It should explain in brief the subject-matter, date or period for which data belong, basis of classification and the area for which data relate.

(3) **Captions** : Caption means the headings of vertical columns. Data in columns are identified by captions. The wording of caption should be as brief as possible.

(4) **Stubs** : These relate to headings of horizontal rows. Generally more vital data with long descriptions are written in rows.

(5) **Ruling and Spacing** : With a view to give a neat, tidy and attractive looking, there should be proper rulings and spacing in a table. Horizontal rulings are rarely used in a table. Vertical rulings in the sides of the table are not needed. The size of the table should be adjusted according to space available. Major and minor items should be adjusted according to space available. Major and minor items should be provided space according to their relative importance. Items should not be jumbled together.

(6) **Averages and Totals** : Averages are usually placed at the bottom of the numbers (which are averaged). The table should contain a separate column for sub-totals of each class and a separate column for total of combined classes. Average if to be shown should follow the totals from which it was calculated.

(7) **Body of the Table** : The presentation of data in a table should be as comprehensive as possible relevant to the object of investigation. Irrelevant matter be avoided. There should be a systematic arrangement of items in the table either alphabetically or geographically or in some chronological order.

(8) **Footnotes** : Proper footnotes are used only if there is need to call attention to some figures or headings which may not be understandable without these notes. Figures not available may be indicated by the words 'N.A.' and negligible figures by a dash (-).

4.12 Essentials of a Good Table

(i) The table should be simple and compact so that it is readily comprehensible. It should be free from all sorts of over-lapping and ambiguities.

(ii) The classification in the table should be so arranged as to focus attention on the main comparisons and exhibit the relationship between various related items and facilitate statistical analysis. it should highlight the relevant and desired information needed for further statistical investigation and emphasize the important points in a compact and concise way. Different modes of lettering (in italics, bold or antique type, capital letters or small letters of the alphabet etc.) may be used to distinguish points of special emphasis.

(iii) A table should be complete and self explanatory. It should have a suitable title, head note (if necessary), captions and stubs and foot note (if necessary). If the data are secondary, the source note should also be given. The use of dash (-) and ditto marks. (,,) should be avoided. Only accepted common, abbreviations should be used.

(iv) A table should have an attractive get up which is appealing to the eye and the mind so that the reader may grasp it without any strain. This necessitates special attention to the size of the table and proper spacing of rows and columns.

(v) Since a statistical table forms the basis for statistical analysis and computation of various statistical measures like averages, dispersion, skewness etc., it should be accurate and free from all sorts of errors. This necessitates checking and re-rechecking of the entries in the table at each stage because even a minor error of tabulation may lead to very fallacious conclusions and misleading interpretations of the results.

(vi) The classification of the data in the table should be in alphabetical, geographical or chronological order or in order of magnitude or importance to facilitate comparisons.

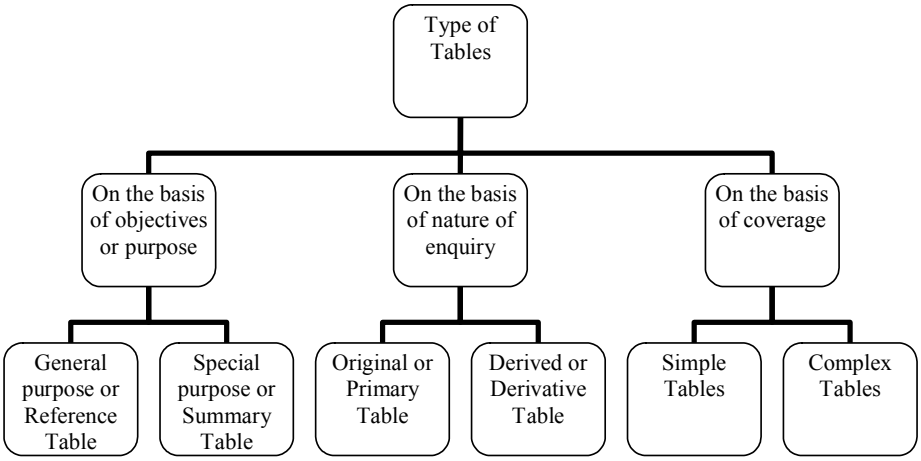
(vii) A summary table should have adequate interpretative figures like totals, ratios, percentages, averages etc.

4.13 Types of Tables

Statistical tables are constructed in many ways. Their choice basically depends upon :

(1) Objectives and scope of the enquiry.
(2) Nature of the enquiry (primary or secondary)
(3) Extent of coverage given in the enquiry.

The following diagrammatic scheme elegantly displays the various forms of tables commonly used in practices.



1. **General Purpose (or Reference) and Special Purpose (or Summary) Tables** :- General purpose tables, which are also known as reference tables or sometimes informative tables provide a convenient way of compiling and presenting a systematically arranged data, usually in chronological order, in a form which is comparative studies, relationship or significance of figures. Most of the tables prepared by government agencies e.g. the detailed tables in the census reports, are of this kind. These tables are of repository nature and mainly designed for use by research workers, statisticians and are generally given at the end of the report in the form of an appendix. Examples of such tables are : age and sex wise distribution of the population of a particular region, community or country; pay rolls of a business house; sales orders for different products manufactured by a concern; the distribution of students in a university according to age, sex and the faculty they join, and so on.

As distinct from the general purpose of reference tables, the special purpose or summary tables (also sometimes called interpretative tables) are of analytical nature and are prepared with the idea of making comparative studies and studying the relationship and the significance of the figures provided by the data. These are generally constructed to emphases some facts or relationships pertaining to a particular or specific purpose. In such tables interpretative figures like ratios, percentages etc., are used in order to facilitate comparisons. Summary tables are sometimes called derived or derivative tables (discussed below) as they are generally derived from the general purpose tables.

2. **Original and Derived Tables** :- On the basis of the nature or originality of the data, the tables may be classified into two classes :

(i) Primary Tables (ii) Derived or Derivative Tables.

In a primary table, the statistical facts are expressed in the original form. It, therefore, contains absolute and actual figures and not rounded numbers or percentages. On the other hand derived or derivative table is one which contains figures and results derived from the original or primary data. It expresses the information in terms of ratios, percentages, aggregates or statistical measures like average, dispersion, skewness etc. For instance, the time series data is expressed in a primary table but a table expressing the trend values and seasonal and cyclic variations is a derived table. In practice, mixtures of primary and derived tables are generally used, an illustration is being given below :

Table No. 4.5

Load Carried by Indian Railway and India Road Transport for Different years

(In Billion Tonnes Km)

Year	Indian Railways	India Road Transport	Percentage Share	
			Railways	Road Transport
2005	88	17	83.8	16.2
2006	117	34	77.5	22.5
2007	125	40	75.8	24.2
2008	122	65	65.2	34.8
2009	134	80	62.6	37.4
2010	148	81	64.6	35.4

3. **Simple and Complex Tables** :- In a simple table the data are classified w.r.t. a single characteristic and accordingly it is also termed as one-way table. On the other hand if the data are grouped into different classes w.r.t. two or more characteristics or criteria simultaneously, then we get a complex or manifold table. In particular, if the data are classified w.r.t. two (three) characteristics simultaneously we get a two-way (three-way) table.

(a) **Simple or Single Table** : A simple table is one in which collected data are presented according to one characteristic only. It is also called as one way table. It is quite easy to study and understand such tables.

Table No. 4.6
Number of Students in Various Faculties of Siddharth College of Commerce & Science

Faculties	Number of Students
Arts	150
Commerce	750
Science	250
Law	1850
Total	**3000**

(b) **Complex Table** : In such tables more than one characteristics of the data are presented. It means the data are divided into more than one category. It may be :

(i) **Double Table** : When data are presented according to two attributes or two characteristics, it is called double table. An example is given below :

Table No. 4.7
Faculty Distribution of Students of Siddharth College of Commerce & Science

Faculties	No. of Students		Total
	Males	Females	
Arts	122	28	150
Commerce	728	22	750
Science	200	50	250
Law	1800	50	1850
Total	**2850**	**150**	**3000**

(ii) **Triple Table** : When data are presented according to three attributes, it is called triple table. The following is an example of such table :

Table No. 4.8
Faculty-wise Distribution of Students of Siddharth College of Commerce & Science
(According to Birth Place)

Faculties	No. of Students								
	Male			Female			Grand Total		
	Urban	Rural	Total	Urban	Rural	Total	Urban	Rural	Total
Arts	100	22	122	20	8	28	120	30	150
Commerce	600	128	728	18	4	22	618	132	750
Science	150	50	200	32	18	50	182	68	250
Law	800	1000	1800	35	15	50	835	1015	1850
Total	**1650**	**1200**	**2850**	**105**	**45**	**150**	**1755**	**1245**	**3000**

4.14 Summary

The process of dividing the collected data into various classes or groups on the basis of their resemblance and similarities is known as classification. The collected data are classified into groups on the basis of their characteristics. This classification can be into groups on qualitative or quantitative basis. Collected data

consists of several characteristics, i.e., there is diversity, and also the data is men for certain purpose, hence there is homogeneity also. The basis of classification is 'in order'. Figures are put in classes or groups in some order, which may be either descending or ascending order.

After classifying statistical data, it becomes necessary to tabulate the same so as to make it simple for comparison and interpretation. Presentation of the classified data in the form of tables so as to make them simple and concise is known as tabulation.

4.15 Key Words

Qualitative Classification - A Classification in which data are classified on the basis of some attributes e.g. intelligence, honesty, etc.

Manifold Classification - In such classification facts are classified in more than one attributes, and then each attribute is further divided into two or more sub-groups.

Quantitative Classification - Such Classification is expressed in numerical figures, and is presented in the form of a statistical series.

Derived Table - The averages, dispersions, correlation etc. calculated from original figures are presented in the form of a table.

4.16 Self Assessment Questions

Q.1 What are the objectives of Classification?

Q.2 Explain the importance of tabulation?

Q.3 Explain the objectives of tabulation?

Q.4 Differentiate between Classification & Tabulation?

Q.5 State the general rules of Tabulation?

Q.6 What are the main parts of a good table?

4.17 Reference Books

1. Garg, Sharma, Jain, Pareek, Business Statistics.

2. Yadav, Jain, Mittal, Statistical Methods.

3. Goyal, Goyal, Jain, Biyani, Gupta, Business Statistics.

4. S.P. Gupta, Statistical Methods.

Unit - 5 Diagrammatic and Graphic Presentation of Data

Structure of Unit:

5.0 Objectives

After completing this unit, you will be able to :

- Give the introduction of diagrammatic and graphic presentation.
- Understand the basic difference between tabulation, diagram and graph.
- Highlight significance of diagrammatic and graphic presentation.
- Through the rules for making diagram and will be able to draft all types of diagrams.
- Draw various graphs to present the data of time series and frequency distribution.
- Enhance the knowledge of interpreting the diagrammatic and graphic presentation.

5.1 Introduction

One of the most important functions of statistics is to remove complexity of quantitative data and make them simple and easily understandable. Tabulation and classification are used for presentation of statistical data in neat, concise, systematic and intelligible form. Another important convincing and easily understood method of presenting the statistical data is diagrammatic presentation. The special feature of diagrammatic presentation is that it attracts attention of common man in dry and boring statistical facts, as they present there facts in the shape of attractive and appealing pictures and charts. Diagrammatic presentation of data is method of visual aids and is very convincing too.

M.J. Moroney – "Cold figures are uninspiring to most people. Diagrams help us to see the pattern and shape of any complex situation. Just as a map gives us a bird's eye-view of the wide stretch of a country, so diagrams help us visualize the whole meaning of a numerical complex at a single glance."

5.2 Difference between Diagrammatic and Tabular Presentation

Diagram and table both are used as tools of presentation of data. Through tabulations and diagrammatic presentation statistical data are presented in very simple and intelligible form but still there are some basic differences in these two methods of data presentation, which are as follows:

1. In tabulation data are presented in form of table while in diagrammatic presentation data are presented in the form of various shapes.

2. In tabulation the grouped data are presented in different columns and rows but in a diagram different shapes are drawn to present the grouped data, such as bars, rectangles, squares, circulars, cubes, etc.

3. Various numbers of grouped data are shown in a table while different shapes are shown in a diagram.

4..Table shows the accurate figure of various groups. On the other hand diagram shows only approximate information.

5. Tables can be used for calculating mode, median and other dividing values, but diagrams can not be used for knowing these values.

6. Table shows accurate value of grouped data so it is very important for advance statistical calculation such as dispersion, skewers, correlation, etc. On the other side diagrams are not useful for any further calculation as the approximate informations are provided by them.

7. Data presented by tabulation seems very boring and unable to attract the common man while the diagrammatic presentation of data automatically seek attention of every common man.

5.3 Significance of Diagrammatic Presentation

Diagrammatic presentation of data is presenting the statistical data in the form of neat and good diagrams and pictures. In this presentation data are presented in various geometrical figures like bars, squares, rectangles, circulars, cubes, etc. The significance of diagrammatic presentation can be assessed by following points:

1. Diagrammatic presentation of data is more attractive then the numerical presentation. In general people ignore numerical figures but they like to see the pictorial figures.

2. Diagrams are visual aids, they present the given numerical data in simple, attractive and appealing form.

3. Diagrams create mere stable effects on the mind of the readers, because a human brain can register pictures for a longer time in comparison to numbers.

4. Diagrams, besides being attractive and fascinating carries another merit and it is that they simplify complexity of quantitative data.

5. Presentation of data through diagrammatic method is helpful for the individual to compose the data and make decisions.

6. An individual needs a lot of time to go through a set of numerical figures but may go through with the data at a glance if presented in a diagram. So it requires less time to make conclusion of data presented by diagrams.

7. For trend analysis diagrams are the most suitable tool to go through because from numericals it is very difficult and time consuming to find out the specific trend in the group of data.

8. Diagrammatic presentation of data is also much useful for the outsiders of a business, who do not want to devote much time in statistical figures but interested in business result and activities.

All the above features make significance of diagrammatic presentation of statistical data in a business and economy.

5.4 Limitations of a Diagram

Diagrams are very attractive and fascinating, they are very easy to understood but they require proper vigilance and intelligence to understand. Some times diagrams can be mis-understood and mis-interpreted as they seem very simple and attractive. Following limitations should be taken care of properly before making use of a diagram:

1. Diagrams are useful for a common man but not for an expert who have to make further research of the data

2. The main limitation is that a diagram always reflects the approximate figures. So it is all right for information purpose. It can not be useful for analysis and decision making.

3. Diagrams can be used only for presentation of approximate values, it is not advisable to use for accurate value presentation.

4. A single diagram is not of much significance. Diagrams can be interpreted only when there is another diagram to compare with.

5. Presently diagrams are misused in the business world. For the purpose of advertisement some facts are shown by diagrams to misguide the consumer.

6. Diagrams are supplement of tabular presentation but they can not replace the table. Multi-faces data can not be presented through diagrams.

5.5 Rules for Making a Diagram and Types of Diagrams

Rules for making a diagram:

1. **Neatness**: - As we say that a diagram is very fascinating and attractive way of data presentation so it is very important that it should be drafted neatly. Diagrams should always made neat, clean and appropriate.

2. **Title**: - Same like tabulation it is also essential for a diagram that it should have a suitable title to indicate the information presented by it. Title of a diagram can be given either at the top of the diagram or at the bottom.

3. **Proportion of length and breadth**: - An appropriate proportion between the length and breadth of a diagram should be maintained. Though there are no set rules are laid down, but in the book "Graphic Presentation" Lutz has suggested a rule called "root two", which means the ratio 1 to 2 or 1 to 1.414 between the smaller side and larger side respectively.

4. **Scale selection**: - The scale of a diagram should be selected in consideration with the size of data to be presented. The size should be appropriate enough to focus the silent features and important characteristics of the grouped data. Normally the scale should be in even numbers or multiples of 5 or 10. The scale must be mentioned very clear and legible and shown separately at both the sides.

5. **Footnotes**: - If some clarification is required then footnotes may be given at the left side bottom of the diagram to explain specific points.

6. **Source notes**: - For the authenticity of a diagram it is necessary to indicate the source from where data have been obtained. At the bottom of the diagram a source note is mentioned for the information of users of diagram.

7. **Index**: - A brief index explaining various types of colours, lines, shapes and designs used in the construction of the diagram should be given for understanding of the users.

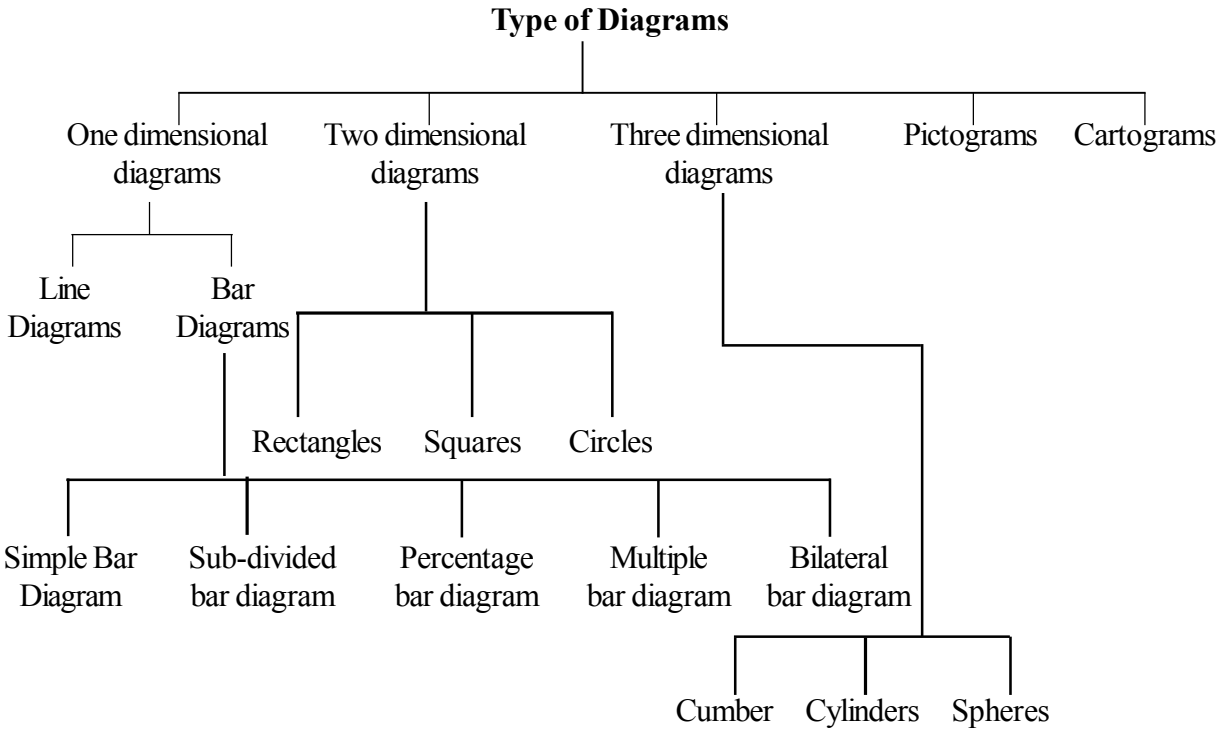
8. **Choice of a Diagram**: - There are many types of diagrams which can be used for presentation of data. Now it is very important to select a diagram, selection of diagram depends upon the size and nature of data as well as the users of diagrams. The important point should be taken care is that right and clear impression of the data should be reflected for interpretations. Hence the choice of a diagram should be done with great caution. All types of diagrams are not suitable for all types of data.

9. **Simplicity**: - Diagrams presents data in a simple way in comparison with tables. Diagrams should be as simple as possible. Do not make them complete, if many informations to be presented, then more than one diagram should be prepared. This will easily understood by a common man who does not have statistical background.

Types of diagrams:

There are variety of diagrams used for data presentation. Most commonly used diagrams can be broadly classified at Annexure – A on page no. 7.

Annexure - A



1. **One dimensional diagrams**:

These are the diagrams in which only one dimension of the figure is reflected through lines or bars. Only the height or length of the diagram can be visualized by these diagrams. In diagrams one dimension can be shown by following ways:

A. **Line Diagrams**:- This is the simplest of all the diagrams. The variables are presented on the x-axis with a suitable scale and the relevant frequencies are shown at y-axis with the suitable scale. This diagram is just vertical lines standing on a common base.

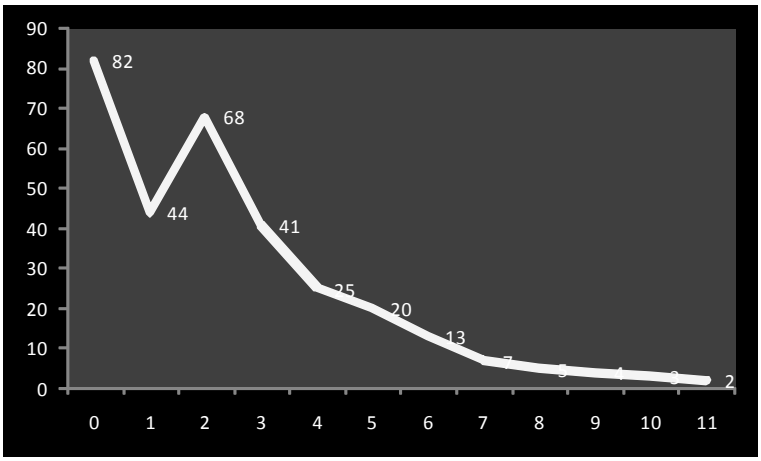
Uses:- Line diagrams are used for comparison and for trend analysis. Time series data can be easily presented by line diagram.

Illustration 1:

The following data shown the number of accidents sustained by 314 drivers of a public utility company over a period of five years.

Number of accidents:	0	1	2	3	4	5	6	7	8	9	10	11
Number of drivers:	82	44	68	41	25	20	13	7	5	4	3	2

Represent the data by a line diagram.



B. **Bar Diagrams**:- Bar in actually a thick line and the height of the bar depends upon the value of frequencies. Bars can be made attractive by colours or shades. Bar diagrams can be of the following types:-

(1) **Simple bar diagrams**:- This is the simplest of the bar diagrams. In simple bar diagrams equal thickness is given to the vertical lines of the line diagram and all bars are drawn on the same base. The space between all the bars is equal and their height depend upon the value of frequency.

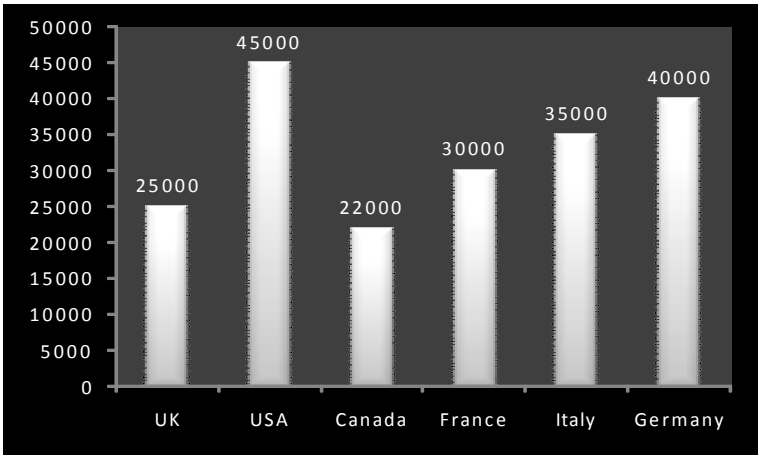
Uses:- Simple bar diagrams are used for the comparative studies.

Illustration 2 :

The following are the number of people visited India during 2009 from various countries. Present them in a simple bar diagram

Country	UK	USA	Canada	France	Italy	Germany
Visitors in ('000)	25	45	22	30	35	40

Solution : Number of visitors from various countries in 2009



(2) **Sub-divided bar diagram:-**

When the total magnitude of the green variables is to be divided into various parts or components, then a sub-divided bar diagram is drawn which is also known as component bar diagram. First of all a bar oftotal is drawn then the bar is divided is various parts according the break up of component values.

Uses:- These diagrams are very useful for comparisons of the totals as well as the different components.

Illustration 3 :

Show the following information by a sub-divided Bar Diagram:

Number of students in a College of Commerce

Year	B.Com.	B.B.A.	B.C.A.	Total
2006-07	1000	700	300	2000
2007-08	1400	800	400	2600
2008-09	1600	1200	600	3400
2009-10	2000	1600	800	4400

Solution: To prepare a sub-divided bar first we make a table of year wise data.

Year wise cumulative data with components

Components	2006-07	2007-08	2008-09	2009-10
B.Com	1000	1400	1600	2000
B.B.A.	700	800	1200	1600
B.C.A.	300	400	600	800
Total	2000	2600	3400	4400

Sub-divided bar diagram of students in a Commerce College

3. **Percentage bar-diagram**:- Many times comparison of data is done on a relative basis. In such cases sub-divided bar diagram can be used, by changing the absolute data into relative data. If the data regarding the cost of production of a particular commodity and its sale price are available for a number of years, sub-divided percentage bars can be drawn for relative comparison.

Uses:- These are specially useful for the study of relative charges in the data

Illustration 4 : Following is the break up of the expenditure of a family on different items of consumption. Draw percentage bar diagram to represent the data.

Item	Expenditure (Rs.)
Food	240
Clothing	66
Rent	125
Fuel and Lighting	57
Education	42
Miscellaneous	190

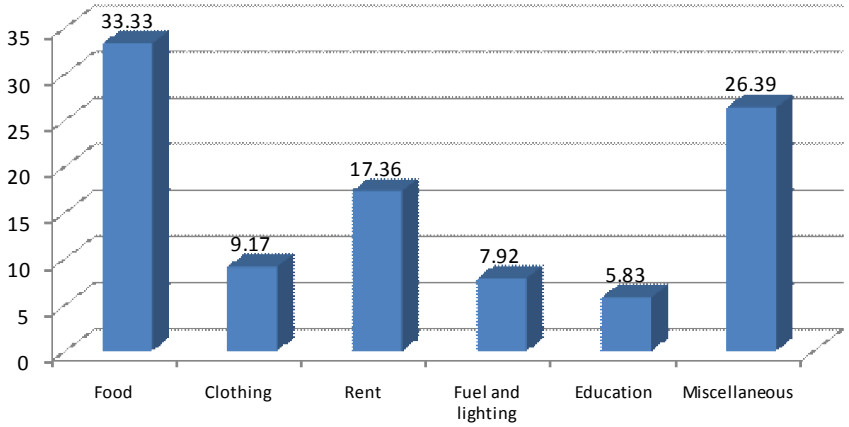
Diagrammatic and Graphic Representation

Solution. First of all we convert the given figures into percentages of the total expenditure as detailed below:

Item	Rs.	Expenditure %	Cumulative %
Food	240	$\frac{240 \times 100}{720} = 33.33$	33.33
Clothing	66	$\frac{66 \times 100}{720} = 9.17$	42.50
Rent	125	$\frac{125 \times 100}{720} = 17.36$	59.86
Fuel and lighting	57	$\frac{57 \times 100}{720} = 7.92$	67.78
Education	42	$\frac{42 \times 100}{720} = 5.83$	73.61
Miscellaneous	190	$\frac{190 \times 100}{720} = 26.39$	100.00
Total	**720**	**100**	

DIAGRAM SHOWING EXPENDITURE OF FAMILY ON DIFFERENT ITEMS OF CONSUMPTION

Percentage Expenditure

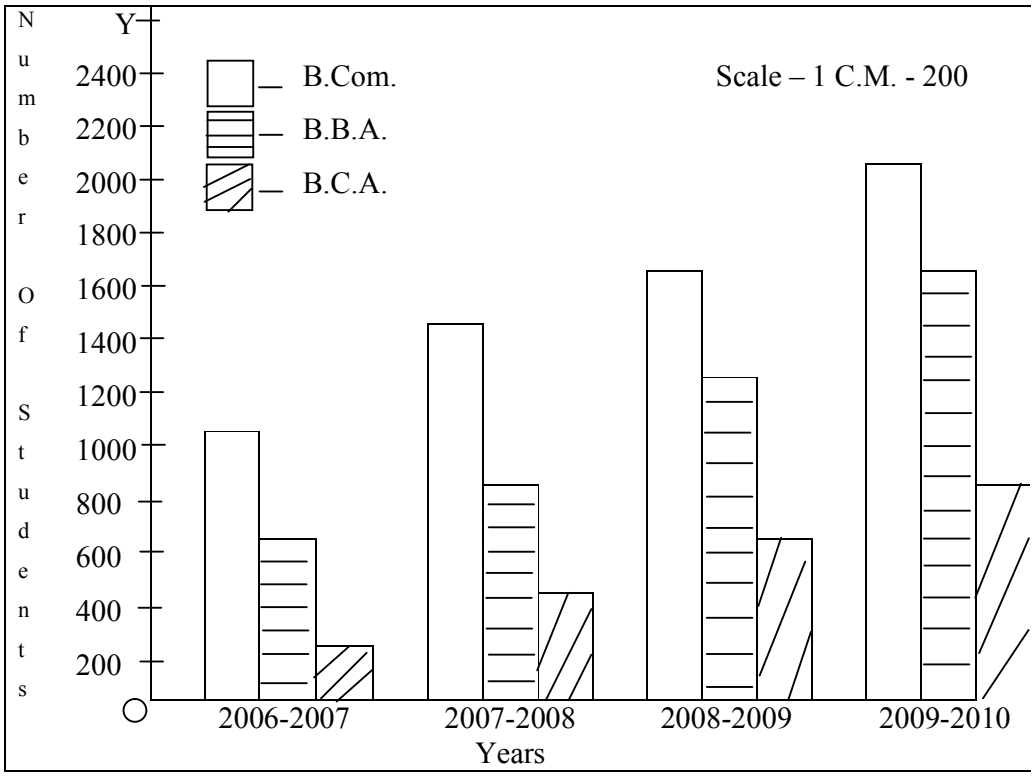


4. **Multiple bar diagram**: The technique of simple bar diagrams can be extended to represent two or more set of inter-related data in one diagram. In this case, a set of adjacent bars (one bar for each variable) is drawn and equal space is given between sets of bars. In a set of adjacent bars different shade or colours are given to different bars and the pattern is followed in other sets of bars.

Uses:- Multiple bar diagram provides information about more than one variable. It also gives the intra comparison of different variable of the same period.

Illustration 5: Present the figures of example 5 in a multiple bar diagram.

Solution: A Triple Bar Diagram showing number of students at various courses in a Commerce College in last years.



5. **Bilateral Bar diagram**:- Bilateral bar diagram which are also known as Deviation bar diagrams. When values are presented with their deviations, which may be positive or negative the Deviation bar diagram is used. In this diagram bars are drawn on both sides of the base either on the left and right or upwards and downwards depending upon the values.

Two Dimensional Diagrams:

These are the diagrams in which length and width both can be drawn at a time. They are also called area diagrams or surface diagrams. Two dimensional diagrams can be simple or sub-divided. They may be in the following shapes: -

(i) Rectangles (ii) Squares, and (iii) Circles

(i) **Rectangles**:- In this diagram the length and width both are considered, because the area of a rectangle is equal to the product of its length and width. Length shows one variable and width shows another depending variable or vice versa.

Just like bars, the rectangles are placed side by side on the same base, proper space is given between different rectangles. It gives a mere detailed information than the bar diagram.

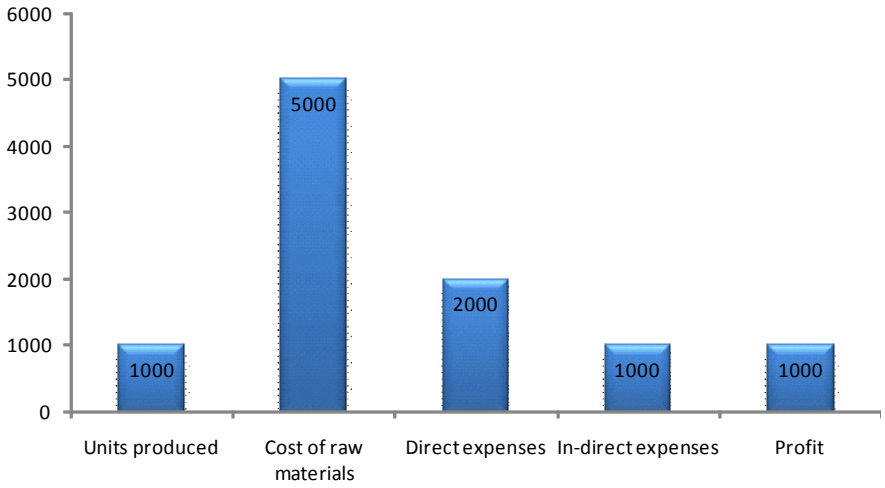
Illustration 6: Prepare a rectangular diagram from the following particulars relating to the production of a commodity in a factory.

Units produced	Rs. 1,000
Cost of raw materials	Rs. 5,000
Direct expenses	Rs. 2,000
Indirect expenses	Rs. 1,000
Profit	Rs. 1,000

Solution: First of all we will find the cost of material, expenses and profits per unit.

DIAGRAM SHOWING COST AND PROFIT FOR A COMMODITY IN A FACTORY

Cost and Profit Per Unit (in Rs.)



(ii) Square is also a two dimensional diagram which reflects two characteristics in a diagram. When some items of the series have values much higher than others, then bar diagram or rectangle is not practical. This kind of situation presented is a square. In the construction of a square first of all the square root of the various figures is calculated and then squares are drawn with the lengths of their sides in the same proportion as the square roots of the original figures. The area of the squares would be in the same proportion of the ratio of original figures.

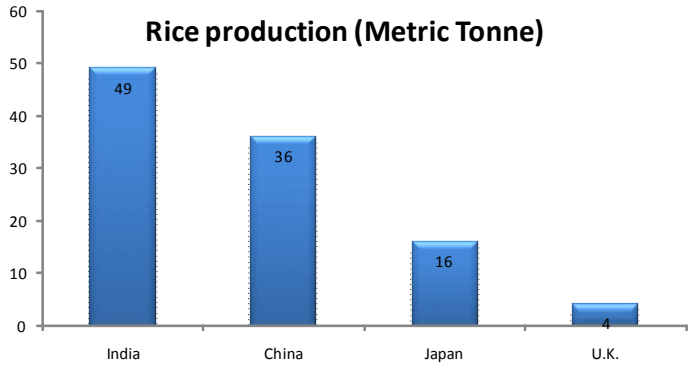
Illustration 7:-Show the followings by square diagrams:

Countries	**Rice Production (in million tones)**
India	49
China	36
Japan	16
U.K.	04

Solution:-

First we will find square roots of the given values.

Indian = 49	=	7	China = 36	=	6
Japan = 16	=	4	U.K. = 4	=	2



(iii) **Circle diagram**:-Circle diagrams are alternative to square diagrams and are used for the same purpose. In circle we have to find out the radius of different circles. The radians are found out by using the following formula – r (pie) $r^2 = 22/7\ r^2$ or $3.1415\ r^2$

To from / frame a circle diagram, first the given values one divide by 22/7 then the root of quotients is calculated and finally the value is divided by equal number to present in form of circle.

Three dimensional diagrams:

Three dimensional diagrams are also known as volume diagrams in which three dimensions, viz. length, width and height are taken into account. They are constructed so that the given magnitude are represented by the volumes of the corresponding diagrams. These diagrams are very useful if there are very wide variations between the smallest and the largest values to be represented. The common forms of such diagrams are cubes, spheres and cylinders.

Pictograms:-

Pictograms is the technique of presenting statistical data through appropriate pictures is one of the very popular devices of diagrammatic presentation. The data are presented through various numbers of pictures or pictures of various sizes as per the magnitude of variables. Pictures are more attractive and appealing to the eye and have a lasting impression on the mind. Accordingly they are extensively used by government and private institutions for diagrammatic presentation of data of public interest.

Pictograms are difficult and time consuming to construct and some how it is not very easy to understand by the common man.

Cartograms:-

It is also known as Map diagram. In this statistical facts are presented through maps along with various diagrams. Cartograms are generally geographical maps where-in different things are presented by different methods. The regional distribution of data is usually shown the use of maps. The distribution of rainfall in various regions of India or the production of wheat in various parts of the country can be shown with help of map. Cartograms are simple and elementary forms of visual presentation and are very easy to understand.

Activity A

A record of daily temperature in the first ten days of April month in the year 2010 is given below. Present the data in a line diagram.

Dates	1	2	3	4	5	6	7	8	9	10
Temperature in Celsius	30	30	32	31	34	32	30	31	33	33

Activity B

Consumption of two commodities in the last three years is given below. Present them by double bar diagram.

Year	Commodity A (in 1000 tons)	Commodity B (in '000 tons)
2009-10	58.0	22.6
2008-09	50.0	20.0
2007-08	46.5	15.5

5.6 Introduction of Graphic Presentation

Graphic presentation is also a visual method of data presentation same as the diagrammatic presentation. Graphic presentation highlights the main features of the collected data, facilitate comparisons among the

data and attracts the common men to take interest in statistical data. From the statistical point of view graphs are better than diagrams. If the relationship between two variables is to be studied diagrams would not be useful for this purpose, graphs can be helpful for this type of study. Graphs are also very useful to study the impact of change in one variable in there is change in another variable by a particular value.

Graphs are the visual aids of presenting the data on a graph paper in the form of lines or curves. This drawing of graphs is also easier than the drawing of diagrams. Graphic presentation shows more accurate numerical value in comparison to diagrams. Graphs save time and labour in the presentation of data. In this presentation data are easily to be understood. They have a better visual impact and easy to graph.

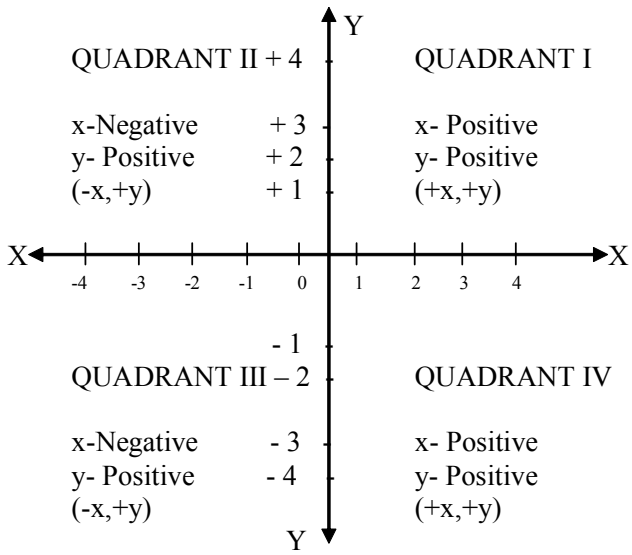
5.7 Advantages of Graphic Presentation

Graphic presentation has a number of advantages that why it is known as the best device of visual presentation of data:

- Graphs are more accurate and precise in comparison to diagrams.
- Being a visual device graphs are attractive, fascinating and appealing than the set of a numerical data.
- Graphic presentation is very useful and easily used for further statistical calculations.
- Positional averages such as medium, mode, etc. can be determined very quickly by a graphic presentation of data.
- The correlation and skewness between two variables is very easy to visualize by a graph. Even their direction and degree of change is also reflected by an accurate graphic presentation.
- Graphs reveal the trends, it is very useful for studying time series and frequency distribution.
- Graphs are drawn on a graph paper having accurate squares, so the lines or curves made on it are very authentic for experts.
- Graphs are easy to draw and less time consuming.

5.8 Construction of a Graph

Graph is always constructed on a specific paper which is known as graph paper. Graph paper has vertical and horizontal lines with equal distance which form equal squares on entire paper. In the construction of graph two simple lines are first drawn which cut each other at right angles. There lines are called axis. The horizontal line is called x-axis or abscissa and the vertical line is called y-axis or ordinate. The point at which they cut each other is called the point of origin. The independent variable is mentioned at x-axis and dependent variable at y-axis. Thus the graph paper is divided into four sections, known as the four quadrants. Generally only the first quadrants is used, the remaining two quadrants are used only when the variables have negative values also.



Along the x-axis the distance measured towards right of the origin or towards right side of the line YOY', are positive and the distances measured towards left of origin or towards left side of the line YOY' are negative. The origin point showing the value zero. Same at y-axis, distance measured towards above the line X'OX one positive and negative at below the lien X'OX.

In the graph of a arithmetic scale, the equal magnitudes of the values of the variables are represented by equal distance along both the axis, though the scale along x-axis and y-axis may be different depending upon nature of the phenomenon under consideration.

5.9 General Rules of Graphic Presentation

To construct a graph will be very easy process if following rules may be kept in mind and the drafted graph will be effective and accurate.

• Title or heading of the graph should be very appropriate and complete. Generally the title of the graph is mentioned at the top of the graph paper.

• Structural frame work of a graph is very important, which describes the position of axes. We have to place the independent variables at x-axis and corresponding values of dependent variables at y-axis, if there are both negative and positive values to be shown then point of origin should be somewhere middle in the graph paper. In case we wish to show only the positive values the point of origin will be at the bottom of left side of graph.

• The scale for both the x-axis and y-axis should be so chosen that the entire data can be accommodated in the available space. The independent variable is shown on x-axis and the dependent variable on y-axis. In plotting of time series the years and months are shown on x-axis and the values y-axis, the scale of x-axis need not begin with O, however the y-axis should have a scale beginning with O. Thus the whole data of time series can be accommodated on a single graph. Proper choice of the scale is necessary for accuracy and attractive presentation.

• Fundamental principal says that scale of y-axis must start from zero, but when the fluctuations in the values of the dependent variables are very high, the vertical scale is stretched by using false base line. In such case the vertical scale is broken and the space between the origin point and the minimum value of the dependent variable is omitted by drawing two zigzag horizontal line above the base line.

• If more than one variable is to be depicted on the same graph, the different graphs so obtained should be distinguished from each other by use of different lines, viz., dotted lines, broken lines, dash-dot lines, thin or thick lines, etc., and an index to identify them should be given.

• Presentation of graph should by first represented by the dots of variables value than these dots are joined by a straight line or curve accordingly.

• Since the main property of graphic presentation is fascinating and attractive so it is very essential that it should be drafted very neatly so that it can print impression for a longer time on viewer's mind. Lines and curves should be drawn very carefully.

• Foot notes regarding some explanations should also be mentioned if necessary. Sources of informations shown in the graph should also be there in source-note for the authenticity of the graph.

• Finally, it should also be taken care that the graph must not be complexed it should be simple or possible.

5.10 Difference between Diagram and Graph

Diagrams and graphs both are visual aids used for presentation of statistical data. Both are attractive, appealing, fascinating and impressive but still there are basic differences between the two, few are as follows:

• Diagrams can be drafted on any paper while graphs can be drafted of graph paper only.

• In diagrams data are presented by lines, bars, rectangles, squares, circles, cubes, etc, on the other side there are only dots, lines and curves can be plotted on a graph.

• In diagrams all lines are always based on a same base viz., on the ox-axis, while in graphs many times false base line is also used.

• Diagrams show only approximate information, which are not much meaningful for the statistical calculation. Graphs are more accurate and precise than the diagrams so quite useful for statistician for the further studies.

• Diagrams are useful in depicting geographical and categorical data, while graphs are used for the study of time series and frequency distributions.

• Diagram construction is quite typical in comparison to construction of graphs.

• Diagrams are not useful for determining the median, mode etc. while graphs can be used for determining all positional averages.

• Graphs can be used for calculation of interpolation, extrapolation and can make a forecast while diagrams are not useful for it.

5.11 Classification of Graphs

Graphs are used mainly for the following types of series

a. Time series b. Frequency distribution.

a. **Time series**:- Graphs showing the frequencies related to time series are called "Historigram" These graphs of time series can be either on natural scale or on ratio scale. If the scale is natural the graphs are "Absolute Historigrams" and when values are converted into index numbers and then presented on the group one "Index Historigrams"

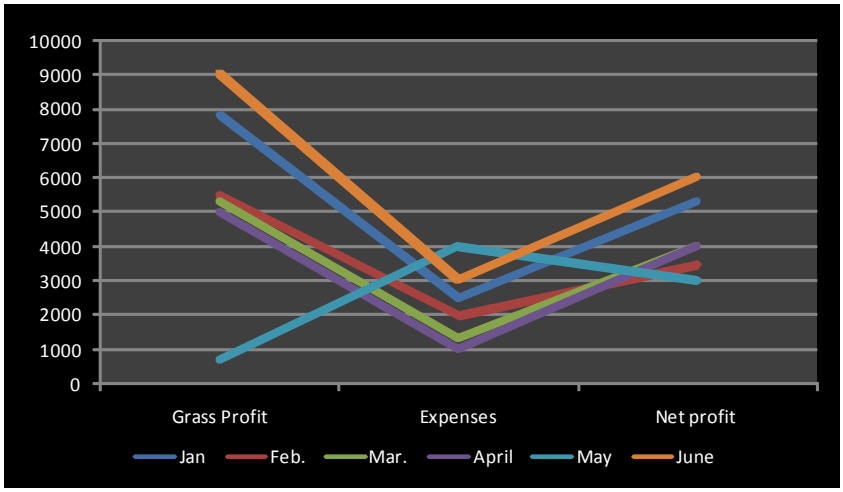
Illustration 8 :

Plot the given data on a graph paper:

Months	**Jan**	**Feb.**	**Mar.**	**April**	**May**	**June**
Grass Profit: (in Rs.)	7800	5500	5300	5000	700	9000
Expenses (in Rs.)	2500	2000	1300	1000	4000	3000
Net profit (in Rs.)	5300	3500	4000	4000	3000	6000

Solution:-

Expenses, Gross profit and net profit of a business

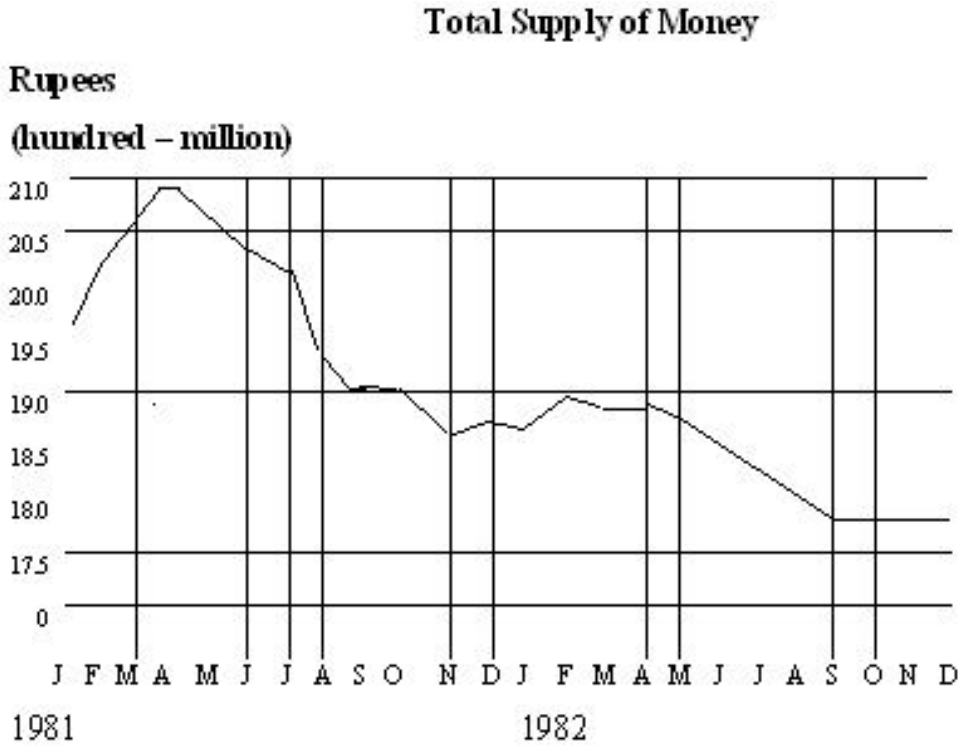


False Base Line

If the fluctuation in the values of a variable are very small as compared to the size of items a false base line is used. By its use even minor fluctuations are magnified so that they are clearly visible on the graph. If the size of items is big and if the vertical scale begins from zero the curve would be mostly on the top of the paper and if the differences in the values of various items are not much, it would, more or less, be of the shape of a straight line.

Total Supply of Money
(in hundred-million rupees)

Month	1981	1982
January	19.7	18.7
February	20.2	19.0
March	20.6	18.9
April	20.9	18.9
May	20.9	18.7
June	20.4	18.5
July	20.1	18.3
August	19.4	18.1
September	19.0	17.9
October	10.0	17.9
November	18.7	17.9
December	18.8	17.9

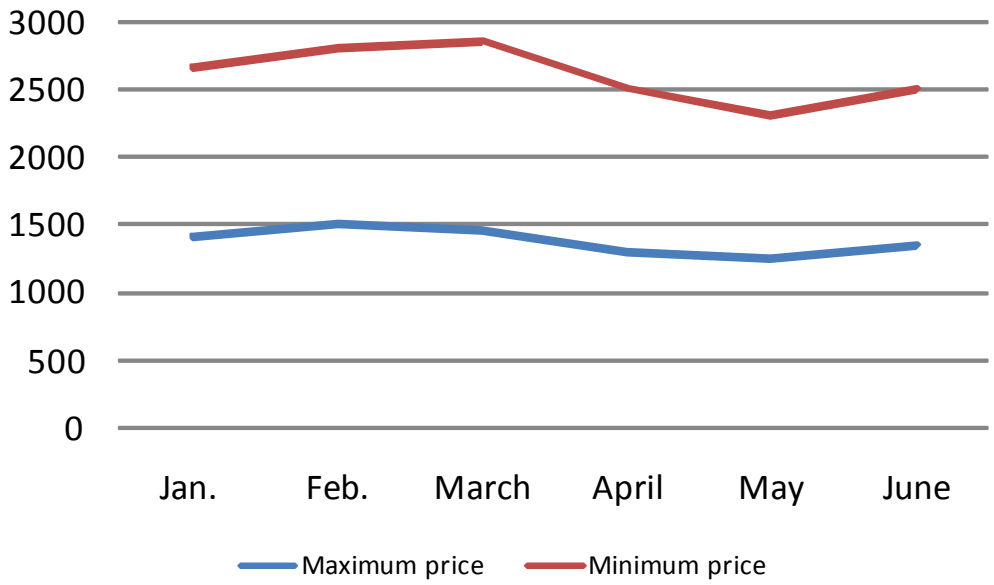


In the above graph 205 centimeters on the vertical scale represents 100 million rupees. If a false base line was not used the vertical scale would have been 52.5 centimeters long. If 2.5 Centimeters was to represent 500 million rupees the size of the vertical scale would still have been more than 11 centimeters but then the fluctuations in the supply of money during these two years would not have very clear from the graph.

Range Charts:- There are graphs which show the range between two variables.

Illustration 9:- The prices of silver in the first half of the year 2010 in Delhi Sarrafa Market are as follows. Plot in a Range Chart.

Month	Jan.	Feb.	March	April	May	June
Maximum price (in Rs.)	1400	1500	1450	1300	1250	1350
Minimum price (in Rs.)	1250	1300	1400	1200	1050	1150



Frequency Distribution:- Graphs of frequency distribution may be of the following types:

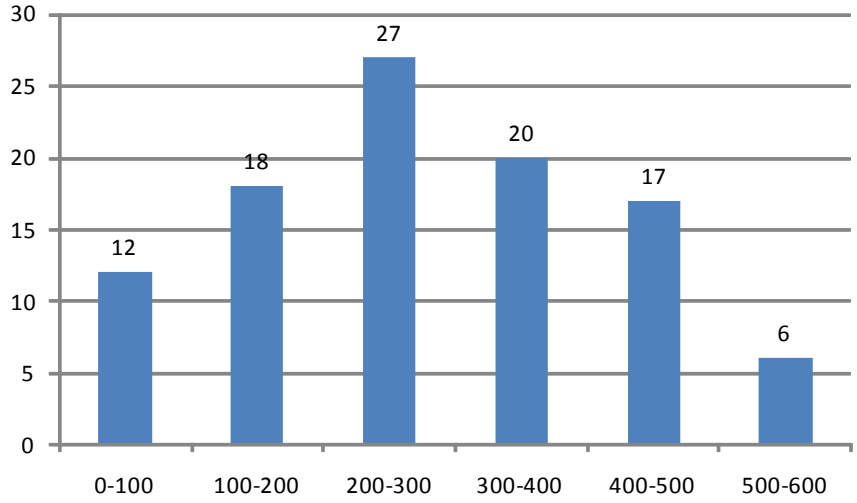
a. Histogram
b. Frequency Polygon
c. Frequency Curve
d. Ogive curve or cumulative frequency curve

a. Histogram:- It is one of the most popular and commonly used devices for showing continuous frequency distribution. The value of variables are taken at x-axis and the frequencies at y-axis.

Illustration 10 : Following are the profit of 100 shops per day:

Profit per Shop	No. of Shops	Profits per Shop	No. of Shops
0-100	12	300-400	20
100-200	18	400-500	17
200-300	27	500-600	6

Determine the 'model value' of the above distribution graphically and verify the result by actual calculation.



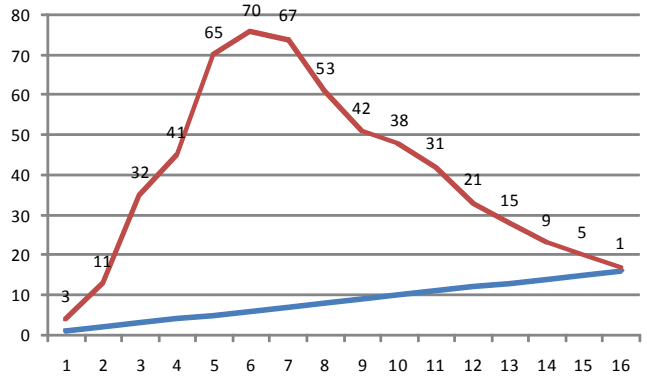
b. **Frequency Polygon**:- Frequency polygon is one more device of graphic presentation of a frequency distribution in all the three series. In case of discrete frequency distribution, frequency polygon is obtained on plotting the frequencies on the vertical axis and joining the points.

In case of grouped or continuous frequency distribution frequency polygon can be drawn by two ways. One is by constructing histogram and another is without constructing the histogram.

Illustration 11:- Values of a variable and their corresponding frequencies:

Values of a Variable	Frequency	Values of a Variable	Frequency
1	3	9	42
2	11	10	38
3	32	11	31
4	41	12	21
5	65	13	15
6	70	14	9
7	67	15	5
8	53	16	2

Values of a Variable and their Corresponding Frequencies



C. Frequency Curve:-

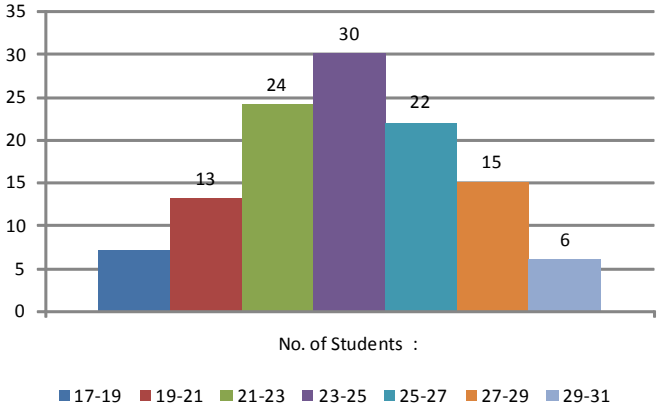
A frequency curve is a smooth free hand curve drawn through the vertices of a frequency polygon. The area covered by the frequency curve is same as that of the histogram of frequency polygon but its shape is smooth one and not with sharp edges.

A frequency curve can be used for interpolation and to give an idea about Skewnell and Kurtosis.

Illustration 12:- Draw a frequency curve for the following distribution.

Age (Yrs)	:	17-19	19-21	21-23	23-25	25-27	27-29	29-31
No. of Students	:	7	13	24	30	22	15	6

Solution.



D. Ogive curve or cumulative frequency curve:- It is a graphic presentation of the cumulative frequency distribution of continuous series.

The frequencies can be cumulated in two ways – by 'less than method' and by 'more than method', the ogive curve will have the shapes accordingly. In less than cumulation the ogive curve will start from the left corner and in more than cumulation it will start from upper side of right corner and end at the base.

Illustration 13:- Prepare 'less than' and 'more than' curves from the weights of the office assistants one graph and find out value of median.

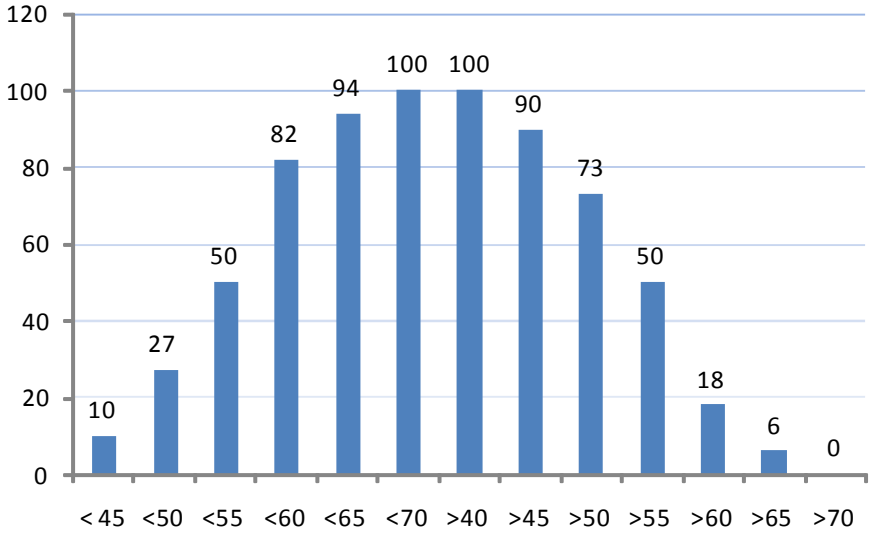
Weight (in kg.)	40-45	45-50	50-55	55-60	60-65	65-70
No. of Office Assistants	10	17	23	32	12	6

Solution

Cumulative Frequency Distribution

Weight (in kg.)	No. of Office Assistants	Weight (in kg.)	No. of Office Assistants
Less than 45	10	More than 40	100
Less than 50	27	More than 45	90
Less than 55	50	More than 50	73
Less than 60	82	More than 55	50
Less than 65	94	More than 60	18
Less than 70	100	More than 65	6
		More than 70	0

Graph Showing Weights of Office Assistants through 'less than and 'more than' Ogive Curves



Weight in Kg.

Activity C :

Draw a Histogram from the following data and find out the value of mode.

Mid value:	5	10	15	20	25	30
Frequency:	8	14	25	22	20	10

Activity D :

Prepare a Histogram and a Frequency Polygon from the data

Class:	0-6	6-12	12-18	18-24	24-30	30-36
Frequency:	4	8	15	20	12	6

5.12 Summary

Diagrammatic and Graphical presentation of data are the grouped statistical data which is very convincing and fascinating. This is the way of presentation through which data and information could reach to the common man. Diagrams make the information attractive and impressive, however, diagrams not so accurate as the tabulated data, even though they are very much capable of conveying the theme. Whenever the data relates to a longer period, graphic presentation yields better results. Diagrammatic presentation is effective for informative data while through graphs element of continuity is maintained and forecasts for the future can easily be made. The very object of data presentation is to provide full set possible information to all concerned. Diagrams and graphs both have their own merits and demerits and, therefore, it is advisable to use both diagrammatic and graphic form of presentation as per the requirement and purpose of users of data.

5.13 Self Assessment Questions

1. What do you mean by diagrammatic and graphic presentation of data ?
2. "Diagrammatic presentation of data is a substitute of tabulation". Do you agree with this statement? Give reasons.
3. Explain the classification of Diagrams.
4. State main differences between diagrammatic presentation and graphic presentation of data.
5. What are the main advantages of graphic presentation?
6. Write short notes on the followings:

   (i) False base line

   (ii) Historigram

   (iii) Histogram

   (iv) Frequency Polygon

   (v) Smoothed Curve

5.14 Reference Books

1. Elhance and Agrawal, Fundamental of statistics.
2. Gupta & Gupta, Business Statistics
3. S.P. Gupta, Practical Statistics.
4. Garg, Sharma, Jain, Pareek, Business Statistics.
5. Yadav, Jain, Mittal, Business Statistics.

Unit - 6 Measures of Central Tendency: Mean, Median and Mode

Structure of Unit:

6.0 Objectives

After completing this unit, you will be able to:

- Understand the measure of central tendency and its uses.
- Comprehend the objectives of measure of central tendency.
- Identify the factors for good measure of central tendency.
- Discuss the relationship of Mean, Mode and Median and its applicability.
- Assess the importance of weight for computing the arithmetic averages.
- Explain the Crude and Standardised death rates.
- Find out the suitable averages in different circumstances.

6.1 Introduction

Central tendency is the middle point of a distribution. Measures of Central tendency are also called measures of location. In order to understand the data properly, these quantitative data are to be classified and converted into a frequency distribution. This process of condensation reduces their bulk and gives prominence to the underlying structure of the data. If the characteristics of the data are properly revealed or if one distribution is to be compared with other, it is necessary that the frequency distribution itself must be summarised and condensed in such a manner that its essence is expressed in as few figures as possible. A single number describing some feature of a frequency distribution is called a descriptive statistic. The statistical analysis would give the single value that describes the characteristics of the entire mass of unwieldy data. This data is called as central value or 'average' or the expected value of the variables. The word average is commonly used in day to day conversation. For example, we often use the word that the average age of the students in the class, average marks in statistics, average income in a group etc. It indicates that the average does not mean that it is very good or very bad but it shows the mediocre type.

6.2 Definitions

The average has been defined by various authors but some of the definitions are given as under:

"An average is sometimes called a 'measure of central tendency' because individual values of the variable usually cluster around it." - **Kellog and Smith**

"Statistics is the science of averages" - **Dr. Bowley**

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

"An average is a typical value in the sense that it is sometimes employed to represent all the individual values in a series or of a variable." **Ya-Lun-Chou**

"An average is a single value selected from a group of values to represent them in some way- a value which is supposed to stand for whole group, of which it is a part, as typical of all the values in the group.

A.E. Waugh

The above definitions give clear-cut indication that the average or measure of central tendency is a single value which represents a group of values. The value has great relevance and significance as it involves the characteristics of whole group.

It is to be connoted that the average represents the entire data, its value lies somewhere in between the two extremes i.e. the largest and the smallest items. This is the reason that the average is frequently referred to as a measure of central tendency.

6.3 Objects

The most important object of calculating and measuring central tendency is to determine a 'single figure' which may be used to represent a whole series involving magnitudes of the same variable. In that sense it is an even more compact description of the statistical data than the frequency distribution. The other object can be enumerated as below:

1. **To present the data in concise form:** The data are represented in the concise form after these go through the classification and tabulation process. This does not provide the meaningful data hence the single value is find out in order to understand it. It can be explained with the help of an example as the marks of all the students of BBA from Vardhaman Mahaveer Open University Kota in a particular subject can not be memorised by all, if any one tries to remember it is of no use. Therefore, the single value by using the measure of central tendency can be obtained which is easy to remember and helpful in drawing the conclusion.

2. **To make Comparison easy:** The comparison of two series of data is done easily by using the measure of central tendency because it provides single value so one can make the comparison with the help of Mean, Mode, Median and so on. Comparison can be made either at a point of time or over a period of time. The comparison of the percentage marks obtained in BBA of the different college, would provide that which college is the best and the pass percentage of the students during the different time period, would give the indication that the result is improved or not. In this way, the comparison is of immense help in framing the suitable and timely policies.

3. **To provide meaningful analysis:** The conclusion can be drawn only when the data are analysed and it depends on the measure of central tendency. Thus, it can be said that the calculation of the measure of central tendency is the basis for analysis.

4. **To act as representative of Universe:** The measure of central tendency provides a single figure which represent the whole of series and it represents the universe from that population it has been taken. Therefore, it is worth mentioning by saying that the measure of central tendency represents the series.

6.4 Requisites of a Measure of Central Tendency

The good Measure of central tendency should include the below enumerated attributes:

1. **Easy to Understand:** It is generally mentioned that the use of statistical method is for simplify the complexity. The Measure of Central should be essentially such that which could easily be understood by the researcher.

2. **Simple in computation:** It is to be connoted that the measure of central tendency should not only be easy to understand but it should be simple in computation work as well so that the acceptability would be more. It has to be kept in mind that the simplicity should not hamper the qualities of a measure of central tendency.

3. **The inclusion of all the items:** The measure of central tendency should be based on the all the items in the series. It should have an impact of change in the value from the series either in the form of increase of decrease. If the measure of central tendency do not bear this attributes than it is not considered to be ideal. For example the arithmetic mean of 20, 30, 40, and 50 is 35. If we drop 20 the average would also change and it would be 40.

4. **Extreme Items should not affect much:** It is worth mentioning that the increase and decrease in the value should have the impact on the measure of central tendency but the extreme item should not have much impact on the value of the measure of central tendency. Otherwise, it would not be regarded as good measure of central tendency. In other words, extremes may distort the average and reduce its usefulness.

5. **Rigid in approach:** The value obtained from the measure of central tendency should have the same interpretation by the different person and the calculation should also be the same.

6. **Capable of algebraic treatment:** The value of the measure of central tendency should involve the mathematical calculation so that the algebraic treatment is possible. It can be cited with the help of an example that the number of student in a class of the department and the average marks in a particular subject of two universities are given, we would be able to compute the combined average of both the universities.

7. **Sampling stability:** The change in the sample from the universe should not have much affect on the measure of central tendency. It means that if we take 10 different groups of university students, and compute the average of each group, we should expect to get the approximately the same value. It may have some difference but it should not be very large.

6.5 Relationship between Mean, Mode and Median

The distribution in which the values of mean, mode and median coincide, it is known as symmetrical distributions means Mean = Median = Mode. In the situation, when these averages i.e. Mean, Mode and Median are not equal, it is called asymmetrical or skewed distribution. In moderately skewed or asymmetrical distributions a very important relationship exists between mean, mode and median. In this case the distance between the mean and the median is about one third the distances between the mean and the mode.

Karl Pearson has given the relationship as enumerated below:

Mode = Mean - 3 (Mean - Median)
Mode = 3 Median - 2 Mean (Z = X- M) and
Median = Mode + 2/3 (Mean - Mode)

If we know the value of any two of the above mentioned averages i.e. Mean, Mode and Median, we can compute the third from these relationship.

Arithmetic Average or Mean

It is the simplest and most easily understandable measure of central tendency. This average is used by the common man in his day to day affairs which means arithmetic mean. Statistician does not use the term average as it is used in the wider perspective. The averages include mean, mode and median as well. Due to it's frequently use it is considered to be same as average but it is a part of the average. It can be computed by adding all the items value and divided by the total number of items.

The average is computed as below:

$$\overline{X} = \frac{\sum X}{N}$$

Where:

$\sum X$ = Total of all value

$\overline{X}$ = Arithmetic Mean

N = Number of items

Illustration 1 :

From the following data, compute the Arithmetic Mean.

Serial No.	1	2	3	4	5
Monthly Income(Rs.)	10000	15000	12000	15000	11000

Solution:

$$\overline{X} = \frac{\sum X}{N} = \frac{63000}{5} = 12600$$

Short-cut Method

It makes the calculation simple and it involves the below mentioned procedures

First arrange the data in ascending or descending order but it is optional and then take the value from assumed mean. Assumed mean may be taken from the given value or it may also be taken from the outside the given value, the effort is to be made to minimise the deviation so that the calculation would become easy.

Secondly the deviation is computed by deducting assumed mean from the given value or size of the item one by one i.e. (X - A)

Thirdly find out the total of all the deviations.

Fourthly use the following formula-

$$\overline{X} = A + \frac{\sum dx}{N}$$

where:

$\overline{X}$ = Arithmetic Mean

A = Assumed Mean

$\sum dx$ = Sum of deviations from Assumed Mean

N = Number of items

Illustration 2 :

Solve the above illustration by short-cut method

Solution :

Serial No.	1	2	3	4	5
Monthly Income (Rs.)	10000	15000	12000	15000	11000
dx (X - 12000)	- 2000	+3000	0	+3000	- 1000

$\sum dx = 3000$

$$\overline{X} = A + \frac{\sum dx}{N} = \overline{X} = 12000 + \frac{3000}{5} = 12000 + 600 = 12600$$

Discrete Series

The calculation in this series can be done according to two methods

(i) *Direct Method:* This method will comprise the following steps :

First, every size or value is multiplied with its frequency.

Second, sum of the multiplied value is obtained.

Third, this sum is divided by total number of frequencies.

Formula is as follows-

$$\overline{X} = \frac{\sum fx}{N}$$

$\overline{X}$ = Arithmetic Mean

$\sum f$ or N = Total of all the frequencies

$\sum fx$ = Sum of the products of value and its respective frequencies.

(ii) *Short-cut Method:* This involves the following procedures.

Firstly take the value as assumed mean.

Secondly the deviation from the all values and assumed mean is to be found out.

Thirdly the product of the value or size and the respective frequencies is computed.

Fourthly the sum of the product is found out

Lastly apply the formula-

$$\overline{X} = A + \frac{\sum fdx}{N}$$

Where:

$\overline{X}$ = Arithmetic Mean

A = Assumed Mean

$\sum fdx$ = Total of the product of deviations and the frequencies.

N = Total of all the frequencies.

Illustration 3 :

From the following data calculate arithmetic mean by direct method and short-cut method.

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

Solution

Calculation of Arithmetic Mean

Direct Method			Short-cut Method			
Marks (X)	f	fX	Marks (X)	f	dx (A=30)	fdx
10	8	80	10	8	- 20	- 160
20	10	200	20	10	- 10	- 100
30	12	360	30	12	0	0
40	20	800	40	20	+10	+ 200
50	6	300	50	6	+ 20	+ 120
60	4	240	60	4	+30	+120
Total	60	1980	Total	60	+180	
$\frac{1980}{60} = 33$					$30 + \frac{180}{60} = 33$	

Hence Arithmetic Mean is 33 marks

Continuous Series

While calculating the arithmetic mean, the series may be inclusive, exclusive of equal classes or unequal classes. The arithmetic means in continuous series is calculated any one of the three methods given below:

(i) Direct Method: In this method the continuous series are converted in a discrete form by finding out the mid value or mid point and the rest of the procedure would remain the same as in direct method of discrete series.

(ii) Short-cut Method: Under this method also the series are converted in the discrete form by finding out the mid values. Here it is assumed that the frequencies are spread evenly within each class over the range of class interval and the rest of the procedure is same as short-cut method of discrete series.

(iii) Step Deviation Method: In this method the deviation is divided by the common value to make the calculation easy and proceed further according to the short-cut method using the following formula-

$$\overline{X} = A + \frac{\sum fdx}{N} Xi$$

Here all the notations would be same as given above except i which stands for step deviation.

Illustration 4 :

From the following data calculate arithmetic mean by direct method, short-cut method and step deviation method.

Wages in Rs.	0-10	10-20	20-30	30-40	40-50	50-60	0-60
No. of Persons	10	15	20	25	18	12	100

Solution :

Calculation of Mean by Direct and Short-cut Method

Wages (Rs.)	f	Mid value (X)	fX	dx (A=25)	fdx
0-10	10	5	50	- 20	- 200
10-20	15	15	225	- 10	- 150
20-30	20	25	500	0	0
30-40	25	35	875	+ 10	+ 250
40-50	18	45	810	+ 20	+ 360
50-60	12	55	660	+ 30	+ 360
Total	100	65	3120		+ 620

$$\overline{X} = A + \frac{\sum fdx}{N} = \frac{3120}{100} = 31.20$$

Hence Arithmetic Mean is 31.20

Calculation of Mean by Step Deviation Method

Wages (Rs.)	f	Mid value (X)	d'x (A=25)	dx/i i = 10	Fd'x
0-10	10	5	- 20	- 2	- 20
10-20	15	15	- 10	- 1	- 15
20-30	20	25	0	0	0
30-40	25	35	+ 10	+ 1	+ 25
40-50	18	45	+ 20	+ 2	+ 36
50-60	12	55	+ 30	+ 3	+ 36
Total	100	65			+ 62

$$\overline{X} = A + \frac{\sum fdx}{N} Xi$$

$$= 25 + \frac{620}{100} X10 = 31.20$$

Hence Arithmetic Mean is 31.20

Mode

The mode is another measure of central tendency. It is the value at the point around which the items are most heavily concentrated. Mode is a positional average and the value of which occurs most frequently in a series. According to Croxton and Cowden, "The mode of the distribution is the value at the points around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values." When one say that the modal wage in a factory is Rs. 150, it indicates that the most of the persons in the factory get Rs. 150 as wages

Calculation of Mode

Individual Series

The mode in this series can not be calculated without converting it into a discrete series as the frequency of every size is one. When we see that the value is repeated many times in the individual series then this series is discrete one not individual one. It should be noted that when the individual series is not converted, the value coming maximum time will be taken as the mode of the series.

Discrete Series

In this series the mode is calculated by the following methods

(i) By Inspection
(ii) By Grouping Method

(i) By Inspection: This method is used when the distribution of frequencies are regular. The regular means when the frequencies first increase and then decreases. In such cases the size having maximum frequencies is known as mode but it is not always true.

As an example, consider the following series:

8, 9, 11, 15, 16, 12, 15, 3, 7, 15

There are ten observations in the series wherein the figure 15 occur maximum number of times-three. The mode is therefore 15.

(ii) By Grouping Method: In the case of distribution of frequencies is irregular, mode is calculated by grouping method. This method will involve the grouping of frequencies and then analysis; this will give the value of mode. For grouping and analysis generally six columns is prepared in a table and below mentioned procedure is taken for computing the mode.

In the first column we write the frequencies and circle the highest frequency.

In the second column we add the two frequencies at a time and circle the highest one.

In the third column we add two frequencies leaving the first frequency at a time and circle the highest one.

In the fourth column we add the three frequencies at a time and circle the highest number (it is to be noted that in the last if two or one frequency is remained leave it)

In the fifth column after leaving the first frequency, add three frequencies at a time and circle the highest one.

In the sixth column leave first two frequencies and add the three frequencies at a time and circle the highest number.

After grouping of the data with the help of circled value of all the columns, we prepare analytical table and find out the mode.

Illustration 5 :

You are required to calculate mode from the following data

Marks	10	15	20	25	30	35	40	42	44	46
No. of Students	4	6	10	15	16	13	17	4	2	1

Solution

Grouping Table

Marks	Frequency					
	1	2	3	4	5	6
10	4	10		20		
15	6		16		31	
20	10	25				**41**
25	15		**31**	**44**		
30	16	**29**			**46**	
35	13		30			34
40	**17**	21		23		
42	4		6		7	
44	2	3				
46	1					

Analysis Table

Column	Size of Items									
	10	15	20	25	30	35	40	42	44	46
1	-	-	-	-	-	-	1	-	-	-
2	-	-	-	-	1	1	-	-	-	-
3	-	-	-	1	1	-	-	-	-	-
4	-	-	-	1	1	1	-	-	-	-
5	-	-	-	-	1	1	1	-	-	-
6	-	-	1	1	1	-	-	-	-	-
Total	-	-	1	3	5	3	2	-	-	-

This is clear from the analysis table that 30 marks come five times is the maximum number of times, therefore, mode is 30.

Continuous Series

In this series first one has to see that the series should be exclusive and class interval should be equal. If series is inclusive one than it is to be converted into exclusive series and if the class interval is unequal, they should be made equal. Then only the modal group will be found either by inspection method or grouping method. After that we use the following formula.

$$Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} Xi$$

Where, L_1 = the lower value of the class in which the mode lies

f_1 = the frequency of the class in which the mode lies

f_0 = the frequency of the class preceding the modal class

f_2 = the frequency of the class succeeding the modal class

i = the class- interval of the modal class

Illustration 6 :

Calculate Mode from the following data

Marks	No. of Students
Less than 10	10
Less than 20	25
Less than 30	37
Less than 40	62
Less than 50	116
Less than 60	171
Less than 70	228
Less than 80	243
Less than 90	248
Less than 100	250

Solution :

Grouping Table

Marks	Frequency					
	1	2	3	4	5	6
0-10	10	25		37		
10-20	15		27		52	
20-30	12	37				91
30-40	25		79	**134**		
40-50	54	**109**			**166**	
50-60	55		**112**			**127**
60-70	**57**	72		77		
70-80	15		20		22	
80-90	5	7				
90-100	2					

Analysis Table

Marks	Frequency					
	1	2	3	4	5	6
0-10	10	25		37		
10-20	15		27		52	
20-30	12	37				91
30-40	25		79	**134**		
40-50	54	**109**			**166**	
50-60	55		**112**			**127**
60-70	**57**	72		77		
70-80	15		20		22	
80-90	5	7				
90-100	2					

The analysis table indicate that the model group is 50-60 which has maximum number as 5. We will use the following formula now:

$$Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} Xi$$

$$Z = 50 + \frac{55 - 54}{(2X55) - 54 - 57} X10$$

$$= 50 + 3.33 = 53.33$$

Hence Mode is 53.33 Marks.

Median

Median is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude. According to Croxton and Cowden, "The median is that value which divides a series so that one half or more of the items are equal to or less than it and one half or more of the items are equal to or greater than it". It is also termed as positional average as it tells us a particular position in a series, that is, the central point.

Calculation of Median

Individual Series

The below enumerated procedure is followed for calculating the mean in this series.
Firstly, the series are to be arranged either in ascending order or descending order.
Secondly, serial numbers is written.
Thirdly, apply the formula which is as under-

$$\text{M} = \text{Size of } \left(\frac{N+1}{2}\right) \text{th item}$$

Where M = size of median; n = number of items

If the series contains odd number then the value of a particular item will be median but where the series have even number of items, then value of the particular item will not be median. It is to be computed by taking the prior item and later item, say if n is 14 then m will come 7.5th item and in order to compute the Median we apply the following formula as well.

$$\text{M} = \frac{\text{Size of } 7^{th} \text{ item} + \text{Size of } 8^{th} \text{ item}}{2}$$

Activity A:

State whether the following statements relating to mean, mode and median are true or false.

1. The value of every observation in the data set is taken into account when we calculate its median.
2. When the population is either negatively or positively skewed, it is often preferable to use the median as the best measure of location because it always lies between the mean and the mode.
3. Measure of central tendency in a data set refer to the extent to which the observations are scattered.
4. With ungrouped data, the mode is most frequently used as the measure of central tendency.
5. If we arrange the observations in a data set from highest to lowest, the data point lying in the middle is the median of the data set.
6. The value most often repeated in a data set is called the arithmetic mean.
7. When working with grouped data, we may compute an approximate mean by assuming that each value in a given class is equal to its midpoint.
8. A mean calculated from grouped data always gives a good estimate of the true value, although it is seldom exact.
9. We can compute a mean for any data set once we are given its frequency distribution.
10. The mode is always found at the highest point of a graph of a data distribution.

For example, we have the following series:

15, 19, 21, 7, 33, 25, 18 and 5

We have to first arrange it in either ascending or descending order. These figures are arranged in an ascending order as follows:

5, 7, 10, 15, 18, 19, 21, 25, 33

Now as the series consists of odd number of items, to find out the value of the middle item, we use the formula-

$$M = \text{Size of } \left(\frac{N+1}{2}\right) \text{th item}$$

Where n is the number of items. In this case, N is 9,

$$M = \text{Size of } \frac{9+1}{2} \text{ th item} = 5, \text{ that is the size of the 5th item is the median}$$

This happens to be 18.

If the above mentioned series consists of one more item 23. We may, therefore, have to include 23 in the above series at an appropriate place, that is, between 21 and 25. Thus, the series is now 5, 7, 10, 15, 18, 19, 21, 23, 25, and 33. Applying the above formula, the median is the size of 5.5th item. Here, we have to take the average of the values of 5th and 6th item. This means an average of 18 and 19, which gives the median as 18.5.

Discrete Series

The following procedure is adopted while computing median in this series

The first step is to arrange the series either in ascending or descending order.

In the second step, find out cumulative frequencies.

Third step is to find out the median by applying formula-

$$M = \text{Size of } \left(\frac{N+1}{2}\right) \text{th item}$$

Where, N stands for total frequencies

The fourth step is to see the M computed in third step in the cumulative frequency. The value in which this cumulative frequency comes would be median.

Where, N stands for total frequencies

The fourth step is to see the M computed in third step in the cumulative frequency. The value in which this cumulative frequency comes would be median.

Illustration 7 :

Compute the median from the following distribution

Height (in inches)	60	61	62	63	64	65	66
No. of Women	27	146	435	398	210	128	98

Solution :

Calculation of Median

Height (in inches)	No. of Women (f)	Cumulative Frequency
60	27	27
61	146	173
62	435	608
63	398	1006
64	210	1216
65	128	1344
66	98	1442

$$M = \text{Size of } \left(\frac{N+1}{2}\right) \text{th item}$$

$$\text{size of } \frac{1442+1}{2} = 721.5 \text{ th item}$$

Size of 721.5th item = this item comes under the cumulative frequency of 1006 for the first time, the size of which is 63. Therefore, median height is 63 inches.

Continuous series :

The procedure for calculating the median under this series is as below:

The series should be in the exclusive continuous form, if it is not there, then it is converted in the exclusive continuous form. The class interval is not made equal as in the case of mode.

Secondly, the cumulative frequencies are made.

Thirdly, median group is to be calculated by applying formula as Size of N/2 th item and seeing this item in the cumulative frequencies, the group concerned will be the median group.

Fourthly, from the median group the median can be calculated by applying the below mentioned formula-

$$M = L_1 + \frac{i}{f}(m - c)$$

Where; M = Median

L_1 = Lower limit of the median group

f = frequency of the median group

i = Class interval of the median group

m = median number

c = cumulative frequency of the group preceding the median group.

If the series is arranged in descending order, then the following formula will be used-

$$M = L_2 + \frac{i}{f}(m - c)$$

L_2 = Upper limit of the median group

Illustration 8 :

You are required to calculate the median from the following information.

Wages (in Rs.)	0-10	10-20	20-30	30-40	40-50
No. of Persons	5	10	12	15	8

Solution :

Calculation of Median

Wages (in Rs.)	No. of Persons	Cumulative frequency
0-10	5	5
10-20	10	15
20-30	12	27
30-40	15	42
40-50	8	50

$$\text{size of } m = \frac{n}{2} \text{ or size of } m = \frac{50}{2} \text{ th item}$$

The 25th item is in the cumulative frequency of 27, the group of which is 20-30. Thus median group is 20-30. By putting the value in the following formula-

$$M = L_1 + \frac{i}{f}(m - c)$$

$$M = 20 + \frac{10}{12}(25 - 15) = 20 + \frac{10X10}{12}$$

$$= 20 + 8.33 = 28.33$$

Curative measures for computing measure of central tendency

S. No.	**Particulars**	**Arithmetic Mean**	**Median**	**Mode**
1	Arrangement of series	Not necessary	Necessary	Necessary
2	Exclusive Series	Not necessary	Necessary	Necessary
3	Inclusive Series	Not to be changed	To be changed	To be changed
4	Open-end classes	Limits to be determined	Not necessary to determine	Not necessary to determine
5	Unequal class interval series	Not to be changed	Change not necessary	Change necessary
6	Mid-value series	Not to be changed	Formation of group essential	Formation of group essential
7	. Cumulative frequency Series	Change necessary	Change necessary	Change necessary

Activity B:

The VP of sales for Vanguard Products has been studying records regarding the performances of his sales reps. He has noticed that in the last two years, the average level of sales per sales rep has remained the same, while the distribution of sales levels has widened. Salespeople's sales levels from this period have significantly larger variations from the mean than in any othe previous two year periods for which he has records. What conclusions you can draw from these observations?

_____

_____

_____

6.6 Weighted Arithmetic Average

The arithmetic mean gives equal importance to all the items but there are certain situations or circumstances when we are required to assign relative importance. Hence, the arithmetic average is not an appropriate measure and at its place weighted average is to be calculated. This can be understood with the help of an example say if we buy the commodities from the market and try to compute arithmetic mean, our conclusion would be wrong as we find dissimilarity in the prices and quantities. In this situation, on the basis of quantities we have to find the weighted average.

Computation of Weighted Arithmetic Average

When weighted arithmetic average is to be calculated, the assignment of weight is essential as these weights may be real or estimated. In case the actual weight is given, it is to be taken but in the absence of this, we may presume the weight. The weight can not be presumed identical by the two people but if it is presumed logically then the conclusions will be the same. The frequencies in the case of discrete and continuous series can be taken as weight for calculating the weighted average arithmetic mean. Here it is important to

note that every frequency can be presumed as weight but every weight can not be presumed as frequency. Only that weight is to be presumed as frequency which has been used at the place of frequency.

The method of computing weighted arithmetic average is as below

(i) Direct Method, (ii) Short-cut Method

(i) Direct Method: Under this method the formula to be used for computing weighted arithmetic average is very similar to the discrete and continuous series of arithmetic mean, the only difference lies is that we use the word weight at the place of frequencies and then find the weighted arithmetic average by multiplying the size or value and weight and divide it by the total of frequencies or weight. The formula is as follows:

$$\overline{X} = \frac{\sum WX}{\sum W}$$

Where:

$\overline{X}$ = Weighted Arithmetic Mean

$\sum WX$ = Sum of the product of the values and weights

$\sum W$ = Total Weight

(ii) Short-cut Method: For computing weighted average arithmetic mean we take one value from the given series as assumed mean and proceed further in the manner similar to discrete series and continuous series of arithmetic mean. The formula is as follow:

$$\overline{X} = A + \frac{\sum Wdx}{\sum W}$$

Where:

A = Assumed Weighted Average Mean

$\sum Wdx$ = Sum of the Product of deviations and weights

$\sum W$ = Total Weight

Illustration 9:

From the following data, you are required to calculate the weighted arithmetic mean and simple arithmetic mean.

Commodities	Quantity Consumed	Price in Rs. Per Kg.
Flour	10 Kg.	2.50
Fuel	50 Kg.	1.50
Sugar	2 Kg	3.50
Oil	3 Kg	4.50

Solution :

The calculation of arithmetic mean do not require the quantity consumed so it is not the part of the computation of arithmetic mean whereas weighted arithmetic mean, the quantity consumed would be taken as weight and this quantity would be multiplied by the price of the respective commodities.

Calculation of Simple Arithmetic Mean and Weighted Arithmetic Mean

Commodities Quantity	Consumed (W)	Price in Rs. per kg. (X)	W x X
Flour	10 kg.	2.50	25.00
Fuel	50 kg.	1.50	75.00
Sugar	2 kg.	3.50	7.00
Oil	3 kg.	4.50	13.50
Total	65 kg.	12.00	120.50

$$\overline{X} = \frac{\sum X}{N} = \frac{12}{4} = 3$$ hence arithmetic mean is 3

$$\overline{X} = A + \frac{\sum Wdx}{\sum W} = \frac{120.50}{65} = 1.85$$ 1.85 hence weighted arithmetic

Activity C:

Give specific example of your own in which:

1. Median would be preferred to Mode.
2. Mode would be preferred to Median.
3. Weighted Arithmetic Mean would be preferred to Arithmetic Mean.
4. Arithmetic mean would be more suitable instead of the Mode and the Median.
5. Median would be most suitable instead of the Arithmetic Mean.

_____

_____

_____

Illustration 10 :

You are required to calculate the weighted arithmetic mean from the following data.

Articles	Index Number	Weight
Wheat	150	30
Rice	120	10
Pulses	140	12
Sugar	180	8
Ghee	160	10

Solution :

Calculation of Weighted Arithmetic Mean

Articles	Index Number	Weight	dx (A=120)	Wdx
Wheat	150	30	+30	+900
Rice	120	10	0	0
Pulses	140	12	+20	+240
Sugar	180	8	+60	+480
Ghee	160	10	+40	+400
Total		70		2020

$$\overline{X} = A + \frac{\sum Wdx}{\sum W}$$

$$120 + \frac{2020}{70} = 120 + 28.86 = 148.86$$

Hence Weighted Arithmetic Mean = 148.86

Utility of Weighted Arithmetic Mean

1. The weighted arithmetic mean is to be used where equal importance is not to be given to all the items. The computation of arithmetic mean in this situation would draw wrong conclusion.

2. The weighted arithmetic mean is useful when the series is distributed among different sub classes.

3. This can be used in the case when percentage and ratios etc. are given.

4. In the case of average salaries of different classes of persons are given and we want to draw conclusions on the basis of all the persons, it is desirable to use weighted arithmetic mean.

5. Generally, weighted arithmetic mean is used for preparation of Index Numbers, Computation of Crude and Standardised death rates, comparison of examination results and calculating average price when we purchase different quantities of different commodities/articles.

Activity D:

You are required to take the population and deaths of your nearest two town under different age groups and suggest that which town is healthier and why?

_____

_____

_____

_____

Activity E:

Suppose that the manager of a hair stylist shop has advertised that 90 percent of the firm's customers are satisfied with the company's services. If a consumer activist believes that this is an exaggerated statement that might require legal action, she can use central tendency techniques to decide whether or not to sue the shop.

_____

_____

_____

_____

6.7 Crude and Standardised Death Rates

The crude and standardised rates are calculated for comparing two places on the basis of population. These rates are generally calculated per thousand.

Crude and General Death Rate: This indicates the number of dead persons in a specific age group of population out of one thousand persons. In this, the death rate of every group is calculated and it is multiplied by weight, i.e. population of the relevant group. The sum of this product is divided by total weights and the quotient is the death rate. It can be calculated by other method as the sum of all the deaths, divided by the number of persons and multiplied by 1000 is death rate per thousand. Formula is given below:

$$\text{Specific Death Rate} = \frac{\text{Death in a specific group}}{\text{Population in the above age group}} X1000$$

$$\text{Crued Death R}ate = \frac{\sum RW}{\sum W}$$

Crude Death rate as per the short-cut method can be enumerated as below:

$$\text{Crude Death Rate or C.D.R} = \frac{\text{Total Deaths}}{TotalPopulation} X1000$$

Standardised Death Rate :

The comparison of the populations of two towns can not be compared on the basis of crude death rates because the number of persons and deaths in different groups differ at two places.

The comparison can not be made until the death rates of standard population are given. Hence, for the purpose of comparison of two towns we presume the population of any one town as standard. The standardised death rate of a town is computed when the specific death rate of that town is multiplied with standardised population, and that product is divided by the total of standard population. This can be understood with the help of an example, say, if we want to calculate the standardised death rate of town S, then we will take the specific death rates of town S and population of other town P which we have presumed as standard. This standardised death rate of 'S' town will be compared with crude death rate of 'P' town because the population of 'P' town have been used in both the cases. By this comparison, it can be said that the town having lower death rate will be considered as a better town as regards health, if, death rate is considered to be an indicator of health. Similarly, birth rates, unemployment rates etc. can also be compared on the same principle.

Illustration 11:

From the following data, find out which town is healthier.

	Town A		Town B	
Age Groups	Population	Death	Population	Death
0-20	5,000	100	10,000	300
20-40	10,000	100	15,000	300
40-60	15,000	225	12,000	120
60-80	12,000	360	8,000	160
80-100	8,000	320	5,000	150

Solution :

Calculation of Crude and Standardised Death Rates

Age Group	Town A				Town B			
	Population W_1	Death	R1 R_1	R1W1	Population W_2	Death	Rate R_2	R2W2
0-20	5,000	100	20	1,00,000	10,000	300	30	3,00,000
20-40	10,000	100	10	1,00,000	15,000	300	20	3,00,000
40-60	15,000	225	15	2,25,000	12,000	120	10	1,20,000
60-80	12,000	360	30	3,60,000	8,000	160	20	1,60,000
80-100	8,000	320	40	3,20,000	5,000	150	30	1,50,000
Total	50,000	1105		11,05,000	50,000	1030		10,30,000

Crude Death Rate of A town

$$\text{C.D.R. of A} = \frac{\sum R_1 W_1}{\sum W_1} = \frac{11,05,000}{50,000} = 22.1 \text{ per thousand.}$$

C.D.R. of A can also be calculated by short-cut method:

$$\text{C.D.R. of A} = \frac{\text{Total Deaths}}{\text{Total Population}} \times 1000 = \frac{1105 \times 1000}{50,000} = 22.1 \text{ per thousand}$$

Crude Death Rate of B town

$$\text{C.D.R. of B} = \frac{\sum R_2 W_2}{\sum W_2} = \frac{10,30,000}{50,000} = 20.6 \text{ per thousand}$$

By short-cut method

$$\text{C.D.R. of B} = \frac{\text{Total Deaths}}{\text{Total Population}} \times 1000 = \frac{1030 \times 1000}{50,000} = 20.6 \text{ per thousand}$$

It can be seen that while comparing the C.D.R.'s of two towns, the death rate of B town is lower as compared to that of A town. This does not mean that the B town is better as in different age groups the population is different. This way the conclusion is wrong from this information. In order to have the proper comparison we have to find out the S.D.R. of one town and C.D.R. of another town. The population in both the town is to be taken the same in such case.

Standardised Death Rate of A town

SDR of A

$$= \frac{\sum R_1 W_2}{\sum W_2} = \frac{(20 \times 10000 + 10 \times 15000 + 15 \times 12000 + 30 \times 8000 + 40 \times 5000)}{(10000 + 15000 + 12000 + 8000 + 5000)}$$

$$= \frac{970000}{50000} = 19.4 \, per 1000$$

It is clear from the above data that while comparing S.D.R of A town and C.D.R. of B town, the S.D.R. of A town is lower, therefore, a town is better than B town as regards health. One more column of R1W2 is to be inserted in the table for the said purpose.When comparison is made with regard to the result of two colleges or the employment condition of two towns, the population of one of the towns will be taken as standard population and only then the conclusions will be drawn.

6.8 Choice of Suitable Averages

When the study of above mentioned averages is done then one may have eagerness to know which average is the best one. It is the natural question which comes in mind and the answer can not be made directly that the particular average is best average and will remain as such under all the circumstances. This is so because each average has its own characteristics. It is seen that in a particular situation one average is best one but at the same time in another situation this may not be the best average. We have witnessed that the use of arithmetic mean is very common but it does not mean that all other averages are useless. It is important to consider the below mentioned points while selecting the suitable averages:

1. **Purpose of Investigation:** The average should be such that which serves the purpose of investigation and it is decided beforehand so that the collection of data is done accordingly.

2. **Nature of data:** The nature of data also affects the type of average to be chosen. For example when one wants to give equal importance to all the values then arithmetic mean is to be good option. If the data are pertaining to honesty, ability etc. we will select the median. The mode is used when the data are connected with the sales and purchases of a business. In this way, it can be connoted that the nature of data affects the selection of a particular average greatly.

3. **Type of data Available:** Frequency distribution of a series also affects the selection of an average. Whether the data is badly skewed (avoid the mean), gappy around the middle (avoid the median), or unequal in class interval (avoid the mode).

4. **Characteristics of an Average:** While selecting a suitable average we should also consider that it should possess the characteristics of an ideal average.

It should further be noted that arithmetic mean takes all the items into consideration, while median and mode are unaffected by extreme values. In many situations, this makes arithmetic mean less attractive. Thus, if we calculate average income for the purposes of determining the general well being of a population, the astronomical incomes of a few individuals give a wrong tilt to the arithmetic mean income. In such a situation, the median income will serve the purpose a lot better.

6.9 Summary

As it has been described that the measure of central tendency should be the best in that particular situation. The summary of the different methods of the measure of central tendency can be given as below:

Arithmetic mean is good to use for studying the economic, social and business problems. This can be used to find out the central tendency of production, income, import, export, price etc. The arithmetic mean should not be used when there is highly skewed distributions, the distribution have open end intervals, when the distribution is unevenly spread, concentration being smaller or large at irregular points and to average ratios and rates of change.

Mode is used to describe qualitative data. This can be used in problems involving the expression of preferences where quantitative measurements are not possible. if we want to compare consumer preferences for different kinds of products, or different kinds of advertising, we can compare the modal preferences expressed by different groups of people but we cannot calculate the median or mean. The best use of average is in a discrete series. The mode is best suited where there is an outstandingly large frequency.

Median is the best average of open-end continuous series because the extreme value does not have the affect on it and average will be representative of the series. The median is also used to study the qualitative facts such as honesty, ability etc. If the income distribution is given, median is the best average.

Weighted arithmetic Mean will give the importance to different items in a series. If we are interested to know the combined average of different groups we will use this average. For calculating percentages, rates etc. we use this average.

6.10 Key Words

Measure of Central Tendency: A measure indicating the value to be expected of a typical or middle data point.

Mean: A central tendency measure representing the arithmetic average of a set of observations.

Mode: The value most often repeated in the data set. It is represented by the highest point in the distribution curve of a data set.

Median: The middle point of a data set, a measure of location that divides the data set into halves.

Median Class: The class in a frequency distribution that contains the median value for a data set.

Symmetrical : A characteristic of a distribution in which each half is the mirror image of the half.

Weighted Mean: An average calculated to take into account the importance of each value to the overall total, that is, an average in which each observation value is weighted by some index of its importance.

6.11 Self Assessment Questions

Q 1. What considerations will weigh with you in choosing a suitable average for studying a phenomenon? Give a few typical cases in which your choice will fall on any average other than the arithmetic average.

Q 2. Define Mode and state the empirical relationship between Mode, Median and Arithmetic Mean of a frequency distribution.

Q 3. What do you mean by Weighted Mean? How does it differ from an un-weighted mean in statistics describing the cases in which the weighted mean is better than the un-weighted mean?

Q 4. Discuss the essential requisites of a measure of a central tendency.

Q 5. Distinguish between Crude Death Rate and Standardised Death Rate.

Q 6. What is meant by 'Central Tendency'? Describe the various measures of measuring the central tendency.

6.12 Reference Books

1. Richard I. Levin and David S. Rubin, Statistics for Management.
2. Gupta, S. P., Statistical Methods.
3. Yadav, Jain, Mittal, Statistical Methods.
4. Nagar, K. N., Statistical Methods.
5. Gupta, C.B. and Gupta, V., An Introduction to Statistical Methods.

Unit - 7 Measures of Dispersion

Structure of Unit:

7.0 Objectives

After completing this unit, you will be able to:

- Understand the concept of dispersion;
- Differentiate between the measures of central tendency and the measures of dispersion;
- State the importance of the measures of dispersion;
- Define quartile deviation and its computation;
- Assess the importance of mean deviation;
- Comprehend standard deviation and its uses to analyse the consistency;
- Using appropriate measure of dispersion in different business situations

7.1 Introduction

In statistics, statistical dispersion (also called statistical variability or variation) is variability or spread in a variable or a probability distribution. Common examples of measures of statistical dispersion are the variance, standard deviation and inter quartile range. Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions. A measure of statistical dispersion is a real number that is zero if all the data are identical, and increases as the data becomes more diverse. It cannot be less than zero. Most measures of dispersion have the same scale as the quantity being measured. In other words, if the measurements have units, such as metres or seconds, the measure of dispersion has the same units.

7.2 Definitions

The measure of dispersion is defined by various experts, some of them are enumerated as below:

"Dispersion is the measure of the variation of the items." **-AL Bowley**

"Dispersion or spread is the degree of scatter or variation of the variable about a central value."

-Brooks &Dick

"The measurement of the scatterness of the mass of figures in a series about an average is called measure of variation or dispersion." **-Simpson & kaska**

"The term dispersion is used to indicate the facts that within a given group, the items differ from one another in size or in other words, there is lack of uniformity in their sizes."**-W. I. King**

"The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data." **- Spiegel**

7.3 Concept of Variation

It is clear from the above that the dispersion measures the extent to which the item vary from some central value. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of second order.

7.4 Methods of Measuring Dispersion

Measures of dispersion may be either absolute or relative. Absolute measures of Dispersion are expressed in same units in which original data is presented but these measures cannot be used to compare the variations between the two series. Relative measures are not expressed in units but it is a pure number. It is the ratios of absolute dispersion to an appropriate average such as co-efficient of Standard Deviation or Co-efficient of Mean Deviation.

Absolute Measures

- Range
- quartile Deviation
- Mean Deviation
- Standard Deviation

Relative Measure

- Co-efficient of Range
- Co-efficient of Quartile Deviation
- Co-efficient of mean Deviation
- Co-efficient of Variation.

Range

The range, in the sense of the difference between the highest and lowest scores, is also called the crude range. When a new scale for measurement is developed, then a potential maximum or minimum will emanate from this scale. This is called the potential (crude) range. Of course this range should not be chosen too small, in order to avoid a ceiling effect. When the measurement is obtained, the resulting smallest or greatest observation will provide the observed (crude) range.

Amongst all the methods of studying dispersion range is the simplest to calculate and to comprehend but its use is rather rare because of the following reasons:

i. Since it is based on the smallest and the largest values of the distribution, it is unduly influenced by two unusual values at either end. On this account, range is usually not used to describe a sample having one or a few unusual values at one or the other end. However, in quality control, range is used as a substitute for standard deviation. With help of a correction factor the value of standard deviation can be known.

ii. It is not affected by the values of various items comprised in the distribution. Thus, it is incapable of giving any information as regards general characters of the distribution within the two extreme observations.

The **range** of a sample (set of data) is simply the maximum possible difference in the data, i.e. the difference between the maximum and the minimum values. A more exact term for it is "**range width"** and is usually denoted by the letter R or w. The two individual values (the max. and min.) are called the **"range limits"**. Often these terms are confused and students should be careful to use the correct terminology.

For example, in a sample with values 2 3 5 7 8 11 12, the range is 10 and the range limits are 2 and 12.

The range is the simplest and most easily understood measure of the dispersion (spread) of a set of data, and though it is very widely used in everyday life, it is too rough for serious statistical work. It is not a "robust" measure, because clearly the chance of finding the maximum and minimum values in a population

depends greatly on the size of the sample we choose to take from it and so its value is likely to vary widely from one sample to another. Furthermore, it is not a satisfactory descriptor of the data because it depends on only two items in the sample and overlooks all the rest. A far better measure of dispersion is the standard deviation (*s*), which takes into account all the data. It is not only more robust and "efficient" than the range, but is also amenable to far greater statistical manipulation. Nevertheless the range is still much used in simple descriptions of data and also in quality control charts.

The Range is calculated as-

Range=L-S

Where:

L=Largest item of the distribution
S=Shortest item of the distribution

For example, let us consider the following three series:

Series: A	6	46	46	46	46	46	46	46
Series: B	6	6	6	6	46	46	46	46
Series: C	6	10	15	25	30	32	40	46

It would be noticed that though in all three series the range is same, i.e. 40, the distributions are not alike: the averages in each case is also quite different. It is because range is sensitive to the values individual items included in the distribution. It thus cannot be depended upon to give any guidance for determining the dispersion of the values within a distribution. Range is extremely sensitive to the size of the sample. As the sample size increases, range also tends to increase, though not proportionately. Thus, the range f the sample is a biased estimate of the variability in the population.

$$\text{Coefficient Range}=\frac{L-S}{L+S}$$

If the averages of the two distributions are the same, a comparison of the range indicates that the distribution with the smaller has less dispersion, and the average of that distribution is more typical of the group.

Illustration 1:

Calculate Coefficient of Range from the following data:

Class Interval Marks

10-20	8
20-30	10
30-40	12
40-50	8
50-60	4

Solution

$$\frac{60-10}{60+10}$$

$$=\frac{50}{70}=\frac{50}{70}=0.714$$

Quartiles and Quartile Range

The quartiles of a data set are formed by the two boundaries on either side of the median, which divide the set into four equal sections. The lowest 25% of the data being found below the first quartile value also called the lower quartile (Q1). The median or second quartile divides the set into two equal sections. The lowest 75% of the data set should be found below the third quartile, also called the upper quartile (Q3). These three numbers are measures of the dispersion of the data, while the mean, median and mode are measures of central tendency.

First quartile (designated Q_1) = lower quartile = cuts off lowest 25% of data = *25th percentile*

Second quartile (designated Q_2) = *median* = cuts data set in half = *50th percentile*

Third quartile (designated Q_3) = upper quartile = cuts off highest 25% of data, or lowest 75% = *75th percentile*

Example-

Given the set {1, 3, 5, 8, 9, 12, 24, 25, 28, 30, 41, 50} we would find the first and third quartiles as follows: There are 12 elements in the set, so 12/4 gives us three elements in each quarter of the set.

So the first or lowest quartile is: **5**, the second quartile is the median **12**, and the third or upper quartile is 28. However some people when faced with a set with an even number of elements (values) still want the true median (or middle value), with an equal number of data values on each side of the median (rather than 12 which has 5 values less than and 6 values greater than). This value is then the average of 12 and 24 resulting in 18 as the true median (which is closer to the mean of 19 2/3. The same process is then applied to the lower and upper quartiles, giving **6.5**, **18**, and **29**. This is only an issue if the data contains an even number of elements with an even number of equally divided sections, or an odd number of elements with an odd number of equally divided sections.

Inter-Quartile Range

The inter quartile range is a statistic which provides information about the spread of a data set, and is calculated by subtracting the first quartile from the third quartile, giving the range of the middle half of the data set, trimming off the lowest and highest quarters. Since the IQR is not affected at all by outliers in the data, it is a more robust measure of dispersion than the range

Inter Quartile Range= $Q_3 - Q_1$

Quartile Deviation (QD) = $\frac{Q_3 - Q_1}{2}$

The quartile deviation is an absolute measure of dispersion. The relative measure corresponding to this measure, called the Coefficient of quartile deviation.

Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Illustration 2: (Individual Series)

Find out the value of quartile deviation and its coefficient from the following data

Roll No.	1	2	3	4	5	6	7
Marks	20	28	40	12	30	15	50

Solution

Computation of QD and Coefficient of QD

Q_1=size of $\frac{N+1}{4}th$ item=size of 7+1/4-2nd item

The size of second item is 15. Thus Q_1=15

Q_3=Size of $3\left(\frac{N+1}{4}\right)$ the item

Size of 6th item is 40. Thus Q_3=40

Coefficient of QD= $\frac{Q_3-Q_1}{Q_3+Q_1}=\frac{40-15}{40+15}=\frac{25}{55}=0.455$

Quartile Deviation or QD= $\frac{Q_3-Q_1}{2}=\frac{40-15}{2}=12.5=\frac{40-15}{2}=12.5$

Illustration 3: (Discrete Series)

Compute coefficient of quartile deviation from the following data:

Marks	10	20	30	40	50	60
No. of Students	4	7	15	8	7	2

Solution

Computation of QD and Coefficient of QD

Marks	10	20	30	40	50	60
No. of Students	4	7	15	8	7	2
c.f.	4	11	26	34	41	43

Q_1=size of $\frac{N+1}{4}th$ item=44 / 4th item = 11th item

The size of 11th item is 20. Thus Q_1=20

Q_3=Size of $3\left(\frac{N+1}{4}\right)$ the item= (3x44)/4th item=33rd item

Size of 33rd item is 40. Thus Q3=40

Quartile Deviation= $\frac{Q_3-Q_1}{2}=\frac{40-20}{2}=10$

oefficient of QD= $=\frac{Q_3-Q_1}{Q_3+Q}=\frac{40-20}{40+20}=0.333$

Illustration 4: (Continuous Series)

Compute coefficient of quartile deviation from the following data:

Marks	Less than 35	35-37	38-40	41-43	Over 43
No. Of Students	14	62	99	18	7

Solution

Computation of QD and Coefficient of QD

Marks	Less than 35	35-37	38-40	41-43	Over 43
No. Of Students	14	62	99	18	7
C.f.	14	76	175	193	200

Q_1=size of $\frac{N}{2}$ th item =size of 200/4=50th item

Q_1 lies in the class 35-37

$$Q_1 = L + \frac{N/4 - c.f.}{f} \text{X} i$$

L=35 N/4=50 c.f. =14 f=62 i=2

$$Q_1 = 35 + \frac{(50) - 14}{62} x2 = 35 + 1.16 = 36.16$$

Q_3=Size of $\frac{3N}{2}$ th item=(3X200)/4=150th item

L=38 3(N/4)=150 c.f.=76 f=99 i=2

$$Q_3 = 38 + \frac{(150) - 76}{99} x2 = 38 + 1.49 = 39.49$$

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2} = \frac{39.49 - 36.16}{2} = 1.67$$

$$\text{Coefficint of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{39.9 - 36.16}{39.9 + 36.16} = \frac{3.33}{75.65} = 0.044$$

Merits or Uses

• It is easiest to calculate and simplest to understand even for a beginner.

• It is one of those measures which are rigidity defined.

• It gives us the total picture of the problem even with a single glance.

• It is used to check the quality of a product for quality control. Range plays an important role in preparing R- charts, thus quality is maintained.

• The idea about the price of Gold and Shares is also made taking care of the range in which prices have moved for the past some periods.

• Meteorological Department also makes forecasts about the weather by keeping range of temp.

Demerits or Limitations or Drawbacks

• Range is not based on all the terms. Only extreme items reflect its size. Hence range cannot be completely representative of the data as all other middle values are ignored.

• Due to above reason range is not a reliable measure of dispersion.

• Range does not change even the least even if all other, in between, terms and variables are changed.

• Range is too much affected by fluctuation of sampling. Range changes from sample to sample. As the size of sample increases range increases and vice versa.

• It does not tell us anything about the variability of other data.

• For open-end intervals, range is indeterminate because lower and upper limits of first and last interval are not given.

Mean Deviation

The mean deviation is also known as the average deviation. it is the average difference between the items in a distribution and the median or mean of that series. Theoretically there is an advantage in taking the deviations from median because the sum of deviations of items from median is minimum when signs are ignored.

However, in practice, the arithmetic mean is more frequently used in calculating the value of average deviation and this is the reason why it is more commonly called mean deviation.

The essence of average deviation lies in the concept of dispersion, which is the average amount of scatter of individual items from either the mean or the median ignoring the algebraic signs. This measure takes into account the whole data. When it is calculated by averaging the deviations of the individual items form their arithmetic mean taking all deviation to be positive, the measures is often called mean deviation.

It may be pointed that we are concerned with the distance of the individual items from their averages and not with their position, which may be above or below the average. The mean deviation is the first measure of dispersion that we will use that actually uses each data value in its computation. It is the mean of the distances between each value and the mean.

Computation of Mean Deviation

Individual Series	**Discrete Series**	**Continuous Series**
$M.D. = \frac{\sum \lvert D \rvert}{N}$	$M.D. = \frac{\sum f \lvert D \rvert}{N}$	$M.D. = \frac{\sum f \lvert D \rvert}{N}$

Where

MD= Mean Deviation

$\sum f \lvert D \rvert$ =sum of multiplication of deviation and frequency

f=frequency

|D|=X-A

X=Observations

A= Assumed Mean

It is the relative measure of mean deviation and obtained by dividing mean deviation by the particular average used in computing mean deviation. So, if mean deviation has been computed from median, the coefficient of mean deviation shall be obtained by dividing mean deviation by median.

Activity A

Make your choices only on the basis of the variability of the distribution. Briefly state the reason for each choice.

(i) The number of points scored by each player in a professional basketball league during an 80 game season.

_____

(ii) The salary of each of 100 people working at roughly equivalent jobs in the State Government.

_____

(iii) The grade point average of each of the 15000 students at a major State University.

_____

Coefficient of Mean Deviation:

$$\text{Coefficient of MD} = \frac{MD}{Median}$$

Illustration 5: (Individual Series)

From the following data given below, compute Mean deviation

Income (In Rs.)	Deviation from median 4400=|D|
4000	400
4200	200
4400	0
4600	200
4800	400
N = 5	$\sum$ |D|=1200

$$\text{Mean Deviation} = M.D. = \frac{\sum |D|}{N}$$

$$\text{Median} = \text{size of } \left(\frac{N+1}{2}\right) \text{ th item} = 5+1/2 = 3^{rd} \text{ item}$$

Size of 3rd item is 4400

$$\text{MD} = \frac{1200}{5} = 240$$

Hence the MD is 240.

Illustration 6: (Discrete Series)

From the following data given below, compute Mean deviation-

X	10	11	12	13	14
F	3	12	18	12	3

Solution

Computation of Mean Deviation

X	f	|D|	f|D|	c.f.
10	3	2	6	3
11	12	1	12	15
12	18	0	0	33
13	12	1	12	45
14	3	2	6	48
	N=48		$\sum f\|D\|$ =36	

$$MD = \frac{\sum f|D|}{N}$$

$$\text{Median} = \text{size of } \left(\frac{N+1}{2}\right) \text{ th item} = 48+1/2 = 24.5^{th} \text{ item}$$

Size of 24.5th item is 12,hence Median=12

MD= $\frac{36}{48} = 0.75$

Coefficient of MD= $\frac{0.75}{12} = 0.0625$

Hence MD is 0.0625

Illustration 7: (Continuous Series)

Calculate Mean Deviation from the following data-

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

Solution

Computation of Mean Deviation

Size	***f***	***c.f.***	***m***	***|D|=m-35.2***	***f|D|***
0-10	7	7	5	30.2	211.4
10-20	12	19	15	20.2	242.4
20-30	18	37	25	10.2	183.6
30-40	25	62	35	0.2	5.0
40-50	16	78	45	9.8	156.8
50-60	14	92	55	19.8	277.2
60-70	8	100	65	29.8	238.4
	N=100				$\sum f\|D\|$=1314.8

Median = size of $\left(\frac{N}{2}\right)$ th item = 100/2= 50th item

Hence, Median lies in the class 30-40

$$M = L + \frac{N/2 - c.f.}{f} \text{X} i$$

$$M = 30 + \frac{50-37}{25} \text{X} 10 = 30 + 5.2 = 35.2$$

Mean Deviation= $M.D. = \frac{\sum |D|}{N} = \frac{1314.8}{100} = 13.148$

Hence the MD is 13.148

Activity B

State whether the following statements relating to measure of dispersion are true or false.

1. The difference between the highest and lowest observations in a data set is called the quartile range.
2. The dispersion of a data set gives insight into the reliability of the measure of central tendency.
3. The standard deviation is equal to the square root of the variance.
4. The inter-quartile range is based on only two values taken from the data set.
5. The standard deviation is measured in the same units as the observations in the data set.
6. The variance, like the standard deviation, takes into account every observation in the data set.
7. The coefficient of variation is an absolute measure of dispersion.
8. The measure of dispersion most often used by statisticians is the standard deviation.

9. One of the advantages of dispersion measures is that any statistic that measures absolute variation also measures relative variation.
10. One disadvantage of using the range to measure dispersion is that it ignores the nature of the variations among most of the observations.
11. The variance indicates the average distance of any observation in the data set from the mean.
12. It is possible to measure the range of open-ended distribution.

Merits of Mean Deviation

The outstanding advantage of of the average deviation is its relative simplicity. it is simple to understand and easy to compute. Any one familiar with the concept the average can readily appreciate the meaning of the average. If a situation requires a measure of dispersion that will be presented to the general public or any group not very familiar with in statistics, the average is useful.

• It is based on each and every data of a dataset. Consequently change in the value of any data would change the value of mean deviation.

• MD is less affected by the value of extreme items than the standard deviation.

• Since the deviation are taken from a central value, comparison about formation of different distributions can easily be made.

Limitations

• The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items.

• This method may not give us very accurate results. The reason is that mean deviation gives best results when the deviations are taken from the median. But median is not a satisfactory measure when the degree of variability in the series is very high.

• It is not capable of further algebraic treatment.

• It is rarely used in sociological studies.

Standard Deviation

The term *standard deviation* was first used in writing by Karl Pearson in 1894, following his use of it in lectures. This was as a replacement for earlier alternative names for the same idea: for Illustration Gauss used "mean error". A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data. Note, however, that for measurements with percentage as unit, the standard deviation will have percentage points as unit.

The Standard Deviation is also known as root mean square deviation for the reason that it is the square root of the mean of the squared deviation from the arithmetic mean.

The standard deviation measures the absolute dispersion (or the variability of distribution; the greater amount of dispersion or variability), the greater the standard deviation, for the greater will be the magnitude of the deviations of the values from their mean.

A small standard deviation means a high degree of uniformility of the observation as well homogeneity of a series or dataset. A large standard deviation means just the opposite. Thus, if we have two or more comparable series with identical means, it is the distribution with the smallest standard deviation that has the most representativeness of the mean.

In probability theory and statistics, the **standard deviation** of a statistical population, a data set, or a probability distribution is the square root of its variance. Standard deviation is a widely used measure of the variability or dispersion, being algebraically more tractable though practically less robust than the expected deviation or average absolute deviation.

It shows how much variation there is from the "average" (mean) (or expected/ budgeted value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.

For Illustration, the average height for adult men in the United States is about 70 inches (178 cm), with a standard deviation of around 3 in (8 cm). This means that most men (about 68 percent, assuming a normal distribution) have a height within 3 in (8 cm) of the mean (67–73 in (170–185 cm)) – one standard deviation, whereas almost all men (about 95%) have a height within 6 in (15 cm) of the mean (64–76 in (163–193 cm)) – 2 standard deviations. If the standard deviation were zero, then all men would be exactly 70 in (178 cm) high. If the standard deviation were 20 in (51 cm), then men would have much more variable heights, with a typical range of about 50 to 90 in (127 to 229 cm). Three standard deviations account for 99.7% of the sample population being studied, assuming the distribution is normal (bell-shaped).

Calculation of Standard Deviation-Individual Series

a. by taking actual mean-

$$\sigma = \sqrt{\left(\frac{\Sigma x^2}{N}\right)}$$

where:

σ = Standard Deviation

$x = X - \overline{X}$

N= Total number of observations

Steps of computing standard deviation from actual mean

1. Calculate the actual mean of the series

2. Take the deviations of the items from the mean i.e. $x = X - \overline{X}$

3. Square these deviations and obtain the total

4. Divide Σx^2 by the total number of observations and extract the square-root. This gives us the value of square root. This gives us the value of standard deviation.

b. by taking assumed mean-

$$\sigma = \sqrt{\left(\frac{\Sigma d^2}{N}\right)} - \sqrt{\left(\frac{\Sigma d}{N}\right)^2}$$

where:

$d = X - A$

σ = Standard Deviation N= Total number of observations

Steps of computing standard deviation from assumed mean

1. Take the Deviations of the items from the assumed mean. Denote these observations by "d". Take the total of these deviations.

2. Square these deviations and obtain the total Σd^2

3. Substitute the values of Σd^2, Σd and N in the above formula.

Illustration 8: (Individual Series)

The daily demands of Ceftum Tablets in a hospital of 10 days are as under:

240,260,290, 245, 255, 288, 272, 263, 277, 251

Calculate standard deviation with the help of assumed mean.

Solution

Computation of Standard Deviation

X	$d= x-264$	d^2
240	-24	576
260	-4	16
290	26	676
245	-19	361
255	-9	81
288	24	576
272	8	64
263	-1	1
277	13	169
251	-13	169
∑X=2641	∑d=+1	$\sum d^2$=2689

$$\sigma = \sqrt{\left(\frac{\Sigma d^2}{N}\right)} - \sqrt{\left(\frac{\Sigma d}{N}\right)^2}$$

$$\sigma = \sqrt{\left(\frac{2689}{10}\right)} - \sqrt{\left(\frac{1}{10}\right)^2} = \sqrt{268.9 - 0.01} = 16.398$$

Hence the SD is 16.398

Calculation of Standard deviation –Discrete Series

a. deviation taken from actual mean

$$\sigma = \sqrt{\left(\frac{\Sigma fx^2}{N}\right)}$$

where:

Σfx^2 =sum of frequency and deviation taken by actual mean

N= sum of frequencies

σ = Standard Deviation

b. deviation taken from assumed mean

$$\sigma = \sqrt{\left(\frac{\Sigma fd^2}{N}\right)} - \sqrt{\left(\frac{\Sigma fd}{N}\right)^2}$$

where:

Σfd = sum of multiplication of frequency and deviation

d= deviation from assumed mean = X-A

Steps of computing standard deviation from actual mean

1. Take the deviations of the items from an assumed mean and denote these deviations by d.

2. Multiply these deviations by the respective frequencies and obtain the total Σfd .

3. Obtain the squares of the deviations and obtain d^2 .

4. Multiply the squared deviations by the respective frequencies, and obtain the total Σfd^2 .

5. Substitute the values in the above formula.

Illustration 9: (Discrete Series)

The annual salaries of a group of lecturers are given in the following table:

Salaries (in'000)	45	50	55	60	65	70	75	80
No. Of Lecturers	3	5	8	7	9	7	4	7

Solution

Computation of Standard Deviation :

Salaries(X)	*No. Of Lecturers(f)*	*d =(x-60)/5*	*fd*	fd^2
45	3	-3	-9	27
50	5	-2	-10	20
55	8	-1	-8	8
60	7	0	0	0
65	9	1	9	9
70	7	2	14	28
75	4	3	12	36
80	7	4	28	112
	N=50		*Σfd=36*	$\Sigma fd^2=240$

$$\sigma = \sqrt{\left(\frac{\Sigma d^2}{N}\right)} - \sqrt{\left(\frac{\Sigma d}{N}\right)^2} \text{ X } i = \sqrt{\left(\frac{240}{50}\right)} - \sqrt{\left(\frac{36}{50}\right)^2} X5$$

$$\sqrt{4.8 - 0.51} X5 = 10.35$$

Hence the standard deviation is 10.35

Calculation of Standard Deviation –Continuous Series

$$\sigma = \sqrt{\left(\frac{fd^2}{N}\right)} - \sqrt{\left(\frac{fd}{N}\right)^2} \; Xi$$

where: $d = \frac{(m - A)}{i}$ and *i= class interval*

Illustration 10:

Calculate mean and standard deviation of the following frequency distribution

Marks	*No. of Students*
0-10	5
10-20	12
20-30	30
30-40	45
40-50	50
50-60	37
60-70	21

Solution

Computation of Standard Deviation

Marks	*m*	*f*	*d=(m-35)/10*	*fd*	fd^2
0-10	5	5	-3	-15	45
10-20	15	12	-2	-24	48
20-30	25	30	-1	-30	30
30-40	35	45	0	0	0
40-50	45	50	1	50	50
50-60	55	37	2	74	148
60-70	65	21	3	63	189
		N=200		$\sum fd=118$	$\sum fd^2=510$

$$\overline{X} = \frac{\sum fd}{N} Xi = 35 + \frac{118}{200} X10 = 40.9$$

$$\sigma = \sqrt{\left(\frac{fd^2}{N}\right)} - \sqrt{\left(\frac{fd}{N}\right)^2} Xi$$

$$\sigma = \sqrt{\left(\frac{510}{200}\right)} - \sqrt{\left(\frac{118}{200}\right)^2} X10$$

$$= \sqrt{2.55 - 0.3481} X10$$

$$= 14.839$$

Hence the standard deviation is 14.839

Interpretation and application of Standard Deviation

A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

For example, each of the three populations {0, 0, 14, 14}, {0, 6, 8, 14} and {6, 6, 8, 8} has a mean of 7. Their standard deviations are 7, 5, and 1, respectively. The third population has a much smaller standard deviation than the other two because its values are all close to 7. In a loose sense, the standard deviation tells us how far from the mean the data points tend to be. It will have the same units as the data points themselves. If, for instance, the data set {0, 6, 8, 14} represents the ages of a population of four siblings in years, the standard deviation is 5 years.

As another example, the population {1000, 1006, 1008, 1014} may represent the distances travelled by four athletes, measured in meters. It has a mean of 1007 meters, and a standard deviation of 5 meters.

Standard deviation may serve as a measure of uncertainty. In physical science, for example, the reported standard deviation of a group of repeated measurements should give the precision of those measurements. When deciding whether measurements agree with a theoretical prediction the standard deviation of those measurements is of crucial importance: if the mean of the measurements is too far away from the prediction (with the distance measured in standard deviations), then the theory being tested probably needs to be revised. This makes sense since they fall outside the range of values that could reasonably be expected to occur if the prediction were correct and the standard deviation appropriately quantified. See prediction interval.

Application examples

The practical value of understanding the standard deviation of a set of values is in appreciating how much variation there is from the "average" (mean).

Weather

As a simple example, consider average temperatures for cities. While two cities may each have an average temperature of 15 °C, it's helpful to understand that the range for cities near the coast is smaller than for cities inland, which clarifies that, while the average is similar, the chance for variation is greater inland than near the coast.

So, an average of 15 occurs for one city with highs of 25 °C and lows of 5 °C, and also occurs for another city with highs of 18 and lows of 12. The standard deviation allows us to recognize that the average for the city with the wider variation, and thus a higher standard deviation, will not offer as reliable a prediction of temperature as the city with the smaller variation and lower standard deviation.

Sports

Another way of seeing it is to consider sports teams. In any set of categories, there will be teams that rate highly at some things and poorly at others. Chances are, the teams that lead in the standings will not show such disparity, but will perform well in most categories. The lower the standard deviation of their ratings in each category, the more balanced and consistent they will tend to be. Whereas, teams with a higher standard deviation will be more unpredictable. For example, a team that is consistently bad in most categories will have a low standard deviation. A team that is consistently good in most categories will also have a low standard deviation. However, a team with a high standard deviation might be the type of team that scores a lot (strong offense) but also concedes a lot (weak defense), or, vice versa, that might have a poor offense but compensates by being difficult to score on.

Trying to predict which teams, on any given day, will win, may include looking at the standard deviations of the various team "stats" ratings, in which anomalies can match strengths vs. weaknesses to attempt to understand what factors may prevail as stronger indicators of eventual scoring outcomes.

In racing, a driver is timed on successive laps. A driver with a low standard deviation of lap times is more consistent than a driver with a higher standard deviation. This information can be used to help understand where opportunities might be found to reduce lap times.

Finance

In finance, standard deviation is a representation of the risk associated with a given security (stocks, bonds, property, etc.), or the risk of a portfolio of securities (actively managed mutual funds, index mutual funds, or ETFs). Risk is an important factor in determining how to efficiently manage a portfolio of investments because it determines the variation in returns on the asset and/or portfolio and gives investors a mathematical basis for investment decisions (known as mean-variance optimization). The overall concept of risk is that as it increases, the expected return on the asset will increase as a result of the risk premium earned – in other words, investors should expect a higher return on an investment when said investment carries a higher level of risk, or uncertainty of that return. When evaluating investments, investors should estimate both the expected return and the uncertainty of future returns. Standard deviation provides a quantified estimate of the uncertainty of future returns.

For example, let's assume an investor had to choose between two stocks. Stock A over the last 20 years had an average return of 10%, with a standard deviation of 20 percentage points (pp) and Stock B, over the same period, had average returns of 12%, but a higher standard deviation of 30 pp. On the basis of risk and return, an investor may decide that Stock A is the safer choice, because Stock B's additional 2% points of return is not worth the additional 10 pp standard deviation (greater risk or uncertainty of the expected return). Stock B is likely to fall short of the initial investment (but also to exceed the initial investment) more often than Stock A under the same circumstances, and is estimated to return only 2% more on average. In this example, Stock A is expected to earn about 10%, plus or minus 20 pp (a range of 30% to -10%), about two-thirds of the future year returns. When considering more extreme possible returns or outcomes in future, an investor should expect results of up to 10% plus or minus 60 pp, or a

range from 70% to (“)50%, which includes outcomes for three standard deviations from the average return (about 99.7% of probable returns).

Calculating the average return (or arithmetic mean) of a security over a given period will generate an expected return on the asset. For each period, subtracting the expected return from the actual return results in the variance. Square the variance in each period to find the effect of the result on the overall risk of the asset. The larger the variance in a period, the greater risk the security carries. Taking the average of the squared variances results in the measurement of overall units of risk associated with the asset. Finding the square root of this variance will result in the standard deviation of the investment tool in question.

Activity C

You are the student of a University Department and you are asked by your teacher to test three new kinds of light bulbs. You have three identical rooms to use in the experiment. Bulb 1 has an average lifetime of 1470 hours and variance of 156. Bulb 2 has an average lifetime of 1400 hours and a variance of 81. Bulb 3 has an average lifetime of 1350 hours and a standard deviation of 6 hours. Now rank the bulbs in terms of relative variability. Which was the best bulb?

Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attribute to one of the factors causing variation in their original series. Variance is defined as follows-

$$\text{Variance} = \frac{\sum\left(X - \overline{X}\right)^2}{N}$$

Thus, variance is nothing but the square of the standard deviation-

Variance = σ^2 or

$$\sigma = \sqrt{Variance}$$

Both the variance and standard deviation are measures of variability in population. These two measures are closely related as is clear from the above formula. Variance is the average squared deviation from the arithmetic mean and standard deviation is the square root of the variance.

In probability theory and statistics, the variance is used as one of several descriptors of a distribution. In particular, the variance is one of the moments of a distribution. In that context, it forms part of systematic approach to distinguishing between probability distributions. While other such approaches have been developed, that based on moments has advantages of mathematical and computational simplicity. The variance is a parameter describing a theoretical probability distribution, while a sample of data from such a distribution can be used to construct an estimate of this variance in the simplest cases this estimate can be the sample variance. Variance is non-negative because the squares are positive or zero. The variance of a constant random variable is zero, and the variance of a variable in a data set is 0 if and only if all entries have the same value.

Suppose that the men have a mean body length of 180 and that the variance of their lengths is 100. Suppose that the women have a mean length of 160 and that the variance of their lengths is 50. Then the mean of the variances is (100 + 50) / 2 = 75; the variance of the means is the variance of 180, 160 which is 100. Then, for the total group of men and women combined, the variance of the body lengths will be 75 + 100 = 175. Note that this uses N for the denominator instead of N -1. In a more general case, if the subgroups have unequal sizes, then they must be weighted proportionally to their size in the computations of the means and variances. The formula is also valid with more than two groups, and even if the grouping variable is continuous.

Coefficient of Variation

The standard deviation is the absolute measure of variation. The corresponding relative measure is known as coefficient of variation. this measure developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problems where we want to compare the variability of two or more than two series. That series for which the coefficient of variation is higher is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous. On the other hand, the series for which coefficient of variation is less is said to be less variable or conversely more consistent, more uniform, more stable or more homogeneous.

Coefficient of Variation is denoted by C.V. and is obtained as follows-

Coefficient of Variation or C.V. = $\frac{\sigma}{\overline{X}} X100$

Illustration 11:

Two brands of tyres are tested with the following results:

Life (in' 000 miles)	*No. Of tyre brand*	
	X	*Y*
20-25	1	0
25-30	22	24
30-35	64	76
35-40	10	0
40-45	3	0

1. Which brand of tyres has greater average life?
2. Compare the variability and state which brand of tyres would you use on your fleet of trucks?

Solution

Calculation of CV of Brand X

Life (in' 000 miles)	*m*	*f*	*d=m-32.5/5*	*fd*	fd^2
20-25	22.5	1	-2	-2	2
25-30	27.5	22	-1	-22	44
30-35	32.5	64	0	0	0
35-40	37.5	10	1	10	10
40-45	42.5	3	2	6	12
		Σ N=100		Σ *fd=-8*	Σ fd^2=48

$$\overline{X} = A + \frac{\sum fd}{N} Xi$$

$$\overline{X} = 32.5 - \frac{8}{100} X5$$

$$= 32.5 - 0.4 = 32.1$$

$$\sigma = \sqrt{\left(\frac{fd^2}{N}\right)} - \sqrt{\left(\frac{fd}{N}\right)^2} Xi$$

$$\sigma = \sqrt{\left(\frac{48}{100}\right)} - \sqrt{\left(\frac{-8}{100}\right)^2} X5$$

$$= \sqrt{0.48 - 0.0064} X5$$

$$= 0.6882 X5$$

$$= 3.441$$

$$CV = \frac{\sigma}{\overline{X}} X100$$

$$= \frac{3.441}{32.1} X100$$

$$= 10.72$$

Calculation of CV of Brand Y

Life (in' 000 miles)	*m*	*f*	*d=m-32.5/5*	*fd*	*fd²*
20-25	22.5	0	-2	0	0
25-30	27.5	24	-1	-24	24
30-35	32.5	76	0	0	0
35-40	37.5	0	1	0	10
40-45	42.5	0	2	0	12
		Σ N=100		Σ *fd= -24*	Σ fd^2=24

$$\overline{X} = A + \frac{\sum fd}{N} Xi$$

$$\overline{X} = 32.5 - \frac{24}{100} X5$$

$$= 32.5 - 1.2 = 31.3$$

$$\sigma = \sqrt{\left(\frac{fd^2}{N}\right)} - \sqrt{\left(\frac{fd}{N}\right)^2} Xi$$

$$\sigma = \sqrt{\left(\frac{24}{100}\right)} - \sqrt{\left(\frac{-24}{100}\right)^2} X5$$

$$= \sqrt{0.24 - 0.0576} X5$$

$$= 0.4271 X5$$

$$= 2.136$$

$$CV = \frac{\sigma}{\overline{X}} X100$$

$$= \frac{2.136}{31.3} X100$$

$$= 6.824$$

Answer 1. Since arithmetic mean is more for brand X of tyres, they have greater average life.

Answer 2. Since coefficient of variation is less for brand Y of tyres, they are more consistent and should be preferred for use.

Merits of Standard Deviation

The standard deviation is the best measure of variation because of its mathematical characteristics.

It is based on every item of the data set.

For computing the variability of two or more distributions coefficient of variation is considered to be most appropriate and this is based on mean and standard deviation.

Limitations of Standard Deviation

It is affected by the extreme values in the data set and this is the major drawback.

Activity D

You are required to frame your opinion to give the advice to Bassart Electronics which is considering employing one of two training programs. Two groups were trained for the same task. Group I was trained by program A; group II, by program B. For the first group, the times required to train the employees had an average of 32.11 hours and a variance of 68.09. In the second group, the average was 19.75 hours and the variance was 71.14. Which training programme has less relative variability in its performance?

_____

7.5 Summary

Dispersion is the spread of the data in a distribution, that is, the extent to which the observations are scattered. Measures of dispersion are also known as averages of the second order as they are second in number from averages. Dispersion gives additional information that enables to judge the reliability of measure of central tendency. If data are widely dispersed, the central location is less representative of the data as a whole than it would be for data more closely centered on the mean. The dispersion gives indication about the widely dispersed data and in that case we may able to handle the situation.

Financial analysts are concerned about the dispersion of a firm's earnings. Widely dispersed earnings – those varying from extremely high to low or even negative levels – indicate a higher risks to stockholders and creditors than do earnings remaining relatively stable. Similarly, quality control experts analyse the dispersion of a product's quality levels. A drug that is average in purity but ranges from very pure to highly impure may endanger lives. Many powerful analytical tools in statistics such as correlation analysis, the testing of hypothesis, the analysis of variance, the statistical quality control, regression analysis are based on measures of variation of one kind or another.

The measures of dispersion are useful in many situations. The average income in a community is not an adequate index of the well being of a community because it glosses over the inequalities of the distribution of income. The measure of dispersion brings out this inequality. Dispersion is helpful in designing a production control system which is based on the premise that if a process is under control, the variability it produces is same over a period of time. If the scatter produced by a process changes over time, it invariably means that something has gone wrong and needs to be corrected.

7.6 Key Words

Dispersion: The spread or variability in a set of data.

Measure of Dispersion: A measure describing how the observations in a data set are scattered or spread out.

Standard Deviation: The positive square root of the variance; a measure of dispersion in the same units as the original data, rather than in the squared units of the variance.

Variance: A measure of the average squared distance between the mean and each item in the population.

Range: The distance between the highest and lowest values in a data set.

Inter quartile Range: The difference between the values of the first and the third quartiles; this difference indicates the range of the middle half of the data set.

Coefficient of Variation: A relative measure of dispersion, comparable across distributions, that expresses the standard deviation as a percentage of the mean.

7.7 Self Assessment Questions

Q 1 What do you mean by dispersion? How it is defined? Why there is a need to calculate the dispersion?

Q 2 What is standard deviation? Explain the method of its computation, its merits and demerits.

Q 3 How is Range and its coefficient calculated? Discuss its merits and demerits.

Q 4 "Frequency distribution may either differ in the numerical size of their averages though not necessarily in their formation, or they may have the same values of their averages yet, differ in their respective formations." Comment.

Q 5 What do you understand by the term Mean Deviation? How it is calculated? What are its advantages and disadvantages?

Q 6 Give the comparative study of different measures of dispersion.

7.8 Reference Books

1. Richard I. Levin and David S. Rubin, Statistics for Management
2. Gupta, S. P., Statistical Methods
3. Yadav, Jain, Mittal, Statistical Methods.
4. Nagar, K. N., Statistical Methods.
5. Gupta, C.B. and Gupta, Vijay, An Introduction to Statistical Methods.

Unit - 8 Measures of Skewness

Structure of Unit:

8.0 Objectives

After completing this unit, you will be able to :

- Throw light on the meaning of skewness.
- Explain the different tests of skewness.
- Describe the difference between skewness and dispersion
- Distinguish between positive and negative skewness.

8.1 Introduction

In the previous units, you have learnt that the measures of central tendency describe us about the concentration of the observations about the middle of the distribution. The measures of variation depict the idea about the scatter or dispersal of observations from the measure of central tendency. These measures of central tendency and dispersion do not tell us whether the scatter of values on either side of an average is symmetrical or not. If frequencies are arranged in a symmetrical order on both sides of a measure of central tendency, we get a symmetrical distribution. Skewness helps us to study the shape i.e. the symmetry or asymmetry of the distribution.

8.2 Meaning and Definition

Some important definitions of skewness are as follows :

"A distribution is said to be skewed if it is lacking in symmetry, that is, if the measure tend to pile up at one end or the other." ***—Paden and Lindquist***

"Skewness or asymmetry is the attribute of frequency distribution that extends further on one side of the class with the highest frequency than on the other." ***—Simpson & Kafka***

"Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution. This characteristic is of particular importance in connection with judging the typicality of certain measures of central tendency." ***—Morris Humburg***

"When a series is not symmetrical, it is said to be asymmetrical or skewed." ***—Croxton & Cowden***

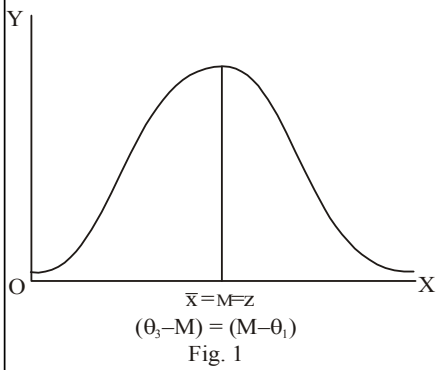
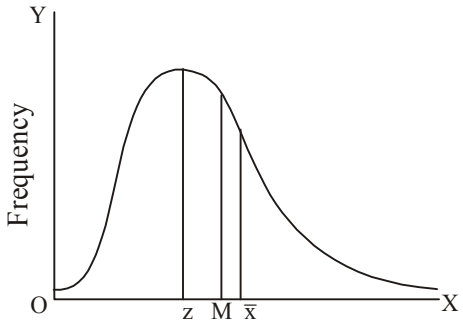
"Skewness is lack of symmetry. When a frequency distribution is plotted on a chart, skewness present in the items tends to disperse chart more on one side of the mean than on the other."

—Riggleman and Frishbee

Thus it is clear that any measure of skewness indicates the difference between the manner in which items are scattered in a particular distribution compared with a normal distribution. There may be two distribution having the same mean and the same standard deviation still their shapes may be quite different. One may be symmetrical and other may be asymmetrical.

8.3 Types of Skewness

The analysis of above definition shows that the term skewness refers to lack of symmetry. If a distribution is normal there would be no skewness in it. The curve drawn from the distribution would be symmetrical. In case of skewed distribution the curve drawn would be tilted either to the left or towards the right. The following three figures give an idea about different types of skewness.

Normal or Symmetrical	**Positive Skewness**
 $(\theta_3 - M) = (M - \theta_1)$ Fig. 1	 $\bar{x} > M > Z$ Fig. 2
Fig. No. 1 gives the shape of an ideal symmetrical curve. It is bell shaped and there is no skewness. The value of mean, median and mode will coincide at one point.	***Fig.No.***3 also reveals moderately skewed distribution. It is tilted towares left. In this case value of mode will be greater than the value of median, and the value of median will be greater than the value of mean. Thus such curve depicts negative skewness.

Negative Skewness

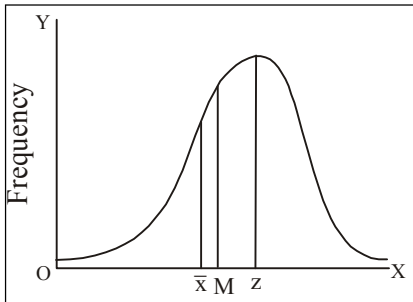
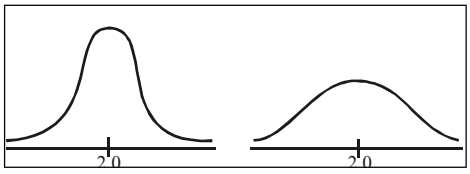


Fig.3

Fig.No.3 also reveals moderately skewed distribution. It is tilted towares left. In this case value of mode will be greater than the value of median, and the value of median will be greater than the value of mean. Thus such curve depicts negative skewness.

Activity A :

A firm using two different methods to ship order to its customers found the following distributions of delivery time for the two methods, based on past records. From available evidence, which shipment, method would you recommend? And why?



8.4 Tests of Skewness

With an object to ascertain whether the distribution is normal or skewed the following tests may be applied.

1. **Relationship between Averages :** If in a distribution the values of Mean, Median and Mode are not identical, it is a skewed distribution. The greater is the difference between Mean and Mode, more will be skewed distribution.

2. **Distance of Pair of quartiles from Median :** If in a distribution the values of Q_3 and Q_1 are equi-distant from median value, it is a symmetrical distribution. If they are not equi-distant, it is a skewed distribution.

3. **Frequencies on either sides of Mode :** If the total frequencies on both sides of the model value are not equal, it is a skewed distribution.

4. **Total of Deviation :** If the sum of positive deviations from the value of Median or Mode are equal to sum of negative deviations, there is no skewness in the distribution.

5. **The Curve :** When the data of a distribution are plotted on a graph paper and if the curve is not bell-shaped (normal), it is a skewed distribution.

8.5 Difference Between Dispersion and Skewness

Basic	Dispersion	Skewness
1. Nature	These measures depicts the scatteredness or spread of values from a measure of central tendency	Measures of skewness show whether the series is symmetrical or asmmetrical. It indicates the shape of the frequency curve.
2. Base	Measure of dispersion depend upon the averages of second order.	Measures of skewness depend upon the averages of first and second orders.
3. Relation with Moments	Measures of dispersion are based upon all the three moments of mean.	Measures of skewness are based on first and third movement only.
4. Conclusion	All the measures of dispersion are positive.	Coefficient of skewness can be positive or negative.
5. Relation with Moments	Measures of dispersion are based upon all the three moments of mean.	Measures of skewness are based on first and third movement only.
6. Normal Distribution	A normal distribution may have some value of dispersion.	Measure of skewness is zero, i.e., there is no skewness in normal distribution.
7. Presentation	Dispersion cannot be presented by means of diagrams.	Skewness can easily be presented by diagrams.

Activity B :

1. Draw the sketch of a skewned frequency distribution and show the position of the mean, median and mode when the distribution is asymmetric.

   _____

2. Distinguish clearly by giving figure between positive and negative skewness.

   _____

3. What is the shape of the distribution when the co-efficient of skewness is zero.

   _____

4. What is the relationship between mean, median and mode in a positively skewed and negatively skewed frequency distribution.

   _____

8.6 Measures of Skewness

Measures of skewness are the devices to find out the direction and the extent of asymmetry in a statistical series.

There are four measures of skewness namely :

1.Karl Pearson's measure
2. Bowley's measure
3.Kelly's measure
4. Moments measure

8.6.1 Karl Pearson's Measure

This measure is based on Mean and Mode. When Mode is ill defined in a distribution then Median is used in place of Mode. To compute the relative measure or coefficient, absolute measure is divided by standard deviation. Symbolically,

(i) Absolute Measure :

$$\text{Skewness } (Sk) = \text{Mean } (\bar{X}) - \text{Mode } (Z)$$

If mode is ill-defined,

$$\text{Skewness } (Sk) = 3\left[\text{Mean } (\bar{X}) - \text{Median } (M)\right]$$

The second formula is based on empirical relationship between averages.

(ii) Relative Measure or Coefficient of Skewness (J)

$$(a)\ J = \frac{\bar{X} - Z}{\sigma} \quad (b)\ J = \frac{3(\bar{X} - M)}{\sigma}$$

where J= Coefficient of skewness, X = Mean, M = Median, Z = Mode, σ = Standard deviation.

Limits : There is no theoretical limit to the measure of Coefficient of Skewness. However, in practice, the values given are not very high and usually lie between $\pm$ 1 in formula (a) and $\pm$ 3 in formula (b).

The direction of skewness is represented by algebraic sign. If it is plus, skewness is positive. Ifit is minus, skewness is negative.

Illutration 1 : The following information relate to two distributions, state which distribution is more skewed:

	Distribution I	**Distribution II**
Mean	200	190
Median	195	195
Standard Deviation	10	10

Solution : The value of Mode is not given in this problem. As such it will be solved by 2nd formula given by Karl Pearson.

Distribution I : $J = \frac{3(\bar{X} - M)}{\sigma} = \frac{3(200-195)}{10} = +1.5$

Distribution II: $J = \frac{3(\bar{X} - M)}{\sigma} = \frac{3(190-195)}{10} = -1.5$

Thus both the distribution reveal the same amount of skewness. But it is positive in I distribution while negative in II distribution.

Illustration 2 : From the following figures, find out the Karl Pearson's Coefficient of Skewness :

Age in Years	10	11	12	13	14	15
No. of Students	5	3	9	6	4	3

Solution :

Calculation of Karl Pearson's Coefficient of Skewness

Age	No. of Students	*dx*	*fdx*	fd^2x
10	5	-3	-15	45
11	3	-2	-6	12
12	9	-1	-9	9
13	6	0	0	0
14	4	+1	4	4
15	3	+2	6	12
Total	**30**	**-**	**20**	**82**

$$\bar{X} = A + \frac{\Sigma fdx}{N} \text{ or } 13 + \frac{-20}{30} \text{ or } 13 - .67 = 12.33$$

Mode is clear by inspection, it is 12

$$\sigma = \frac{1}{N}\sqrt{\Sigma fd^2x - (\Sigma fdx)^2} = \frac{1}{30}\sqrt{82 \times 30 - (-20)^2}$$

or $\frac{1}{30}\sqrt{2460 - 400} = \frac{1}{30}\sqrt{2060} = \frac{1}{30} \times 45.39 = 1.513$

Coeff. of Skewness $(j) = \frac{\bar{X} - Z}{\sigma} = \frac{12.33 - 12}{1.513} = 0.218$

Illustration 3 : Apply Karl Pearson's method and calculate coefficient of skewness from the data given below :

Weight in Kilogram	2.0-2.2	2.2.-2.4	2.4-2.6	2.6-2.8	2.8-3.0
No. of Children	5	28	12	7	3

Solution :

Calculation of Karl Pearson's Coefficient of Skewness

Weight (kg.)	No. of Children (f)	Mid-values (X)	Step dev. A = 2.5/0.2 (dx')	Product of Freq. & dx (fdx')	Product of fdx' & dx' (fd^2x')
2.0-2.2	5	2.1	-2	-10	20
2.2-2.4	28	2.3	-1	-28	28
2.4-2.6	12	2.5	0	0	0
2.6-2.8	7	2.7	+1	+7	7
2.8-3.0	3	2.9	+2	+6	12
Total	**55**	**-**		**-25**	**67**

$$\bar{X} = A + \frac{\Sigma fdx'}{N} \times i = 2.5 + \frac{-25}{55} \times 0.2 \text{ or } 2.5 - 0.09 = 2.41 \text{ kg.}$$

Mode lies in $(2.2 - 2.4)$ group (by inspection). Formula is

$$Z = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 2.2 + \frac{23}{23+16} \times 0.2 \text{ or } 2.2 + 0.12 = 2.32 \text{ kg.}$$

$$\sigma = \frac{i}{N}\sqrt{\Sigma fd^2x'.N - (\Sigma fdx')^2} = \frac{0.2}{55}\sqrt{67 \times 55 - (-25)^2}$$

$$= \frac{0.2}{55}\sqrt{3685 - 625} \text{ or } \frac{0.2}{55} \times 55.32 = 0.20 \text{ kg.}$$

$$j = \frac{\bar{X} - Z}{\sigma} = \frac{2.41 - 2.32}{0.20} = +0.45$$ Thus skewness is positive.

Illustration 4 : Calculate Karl Pearson's Coefficient of Skewness based on median from the following data :

Marks	No. of Students	Marks	No. of Students
above 0	100	above 50	50
above 10	98	above 60	35
above 20	95	above 70	23
above 30	90	above 80	13
above 40	80	above 90	5

Solution : First cumulative frequencies be converted into ordinary frequencies.

Calculation of Mean, Median and S.D.

Marks	Frequency	Mid-value	Step-deviation	fdx'	fd^2x'	Cum. Freq. (cf)
0-10	100-98=2	5	-4	-8	32	2
10-20	98-95=3	15	-3	-9	27	5
20-30	95-90=5	25	-2	-10	20	10
30-40	90-80=10	35	-1	-10	10	20
40-50	80-50=30	45	0	0	0	50
50-60	50-35=15	55	+1	15	15	65
60-70	35-23=12	65	+2	24	48	77
70-80	23-13=10	75	+3	30	90	87
80-90	13-5=8	85	+4	32	128	95
90-100	5-0=5	95	+5	25	125	100
Total	**N = 100**	**-**	**-**	**89**	**495**	

$$\bar{X} = A + \frac{\Sigma fdx'}{N} \times i = 45 + \frac{89}{100} \times 10 = 45 + 8.9 = 53.9 \text{ marks}$$

$$m = \text{Value of } \frac{N}{2} \text{th item} = \text{Value of } \frac{100}{2} \text{th or 50th item}$$

This lies in 50 cum. freq. whose value is in (40 - 50) group

$$M = l_1 + \frac{i}{f}(m - c) = 40 + \frac{10}{30}(50 - 20) \text{ or } 40 + 10 = 50 \text{ marks}$$

$$\sigma = \frac{i}{N}\sqrt{\Sigma fd^2x'N - (fdx')^2} = \frac{10}{100}\sqrt{495 \times 100 - (89)^2}$$

$$= \frac{1}{10}\sqrt{49500 - 7921} = \frac{1}{10}\sqrt{41579} \text{ or } \frac{203.91}{10} = 20.39$$

$$J = \frac{3(\bar{X} - M)}{\sigma} = \frac{3(53.9 - 50)}{20.39} \text{ or } \frac{11.7}{20.39} = 0.574$$

Thus there is adequate degree of positive skewness.

Illustration 5 : Find out Coefficient of variation and Karl Pearson's coefficient of skewness from the following figures :

Income in Rupees	No. of Persons	Income in Rupees	No. of Persons
0 and not exceeding 9	75	40 and not exceeding 49	452
10 and not exceeding 19	100	50 and not exceeding 59	63
20 and not exceeding 29	302	60 and not exceeding 69	25
30 and not exceding 39	603		

Solution :

Calculation of Karl Pearson's Coefficient of Skewness and S.D.

Income Rs.	Frequency	A = 34.5 i = 10 dx'	fdx'	fd^2x'
0-9	75	-3	-225	675
10-19	100	-2	-200	400
20-29	302	-1	-302	302
30-39	603	0	0	0
40-49	452	1	452	452
50-59	63	2	126	252
60-69	25	3	75	225
Total	1,620	-	-74	2,306

$$\bar{X} = A + \frac{\Sigma fdx'}{N} \times i = 34.5 + \frac{-74}{1620} \times 10 \text{ or } 34.5 - .457 = 34.043$$

Mode lies in the group $(29.5 - 39.5)$ (by inspection)

$$Z = l_1 + \frac{\Delta_1}{\Delta_1 - \Delta_1} \times i \qquad \begin{vmatrix} \Delta_1 = 603 - 302 = 301 \\ \Delta_1 = 603 = 452 = 151 \end{vmatrix}$$

$$= 29.5 + \frac{301}{301 + 151} \times 10 = 29.5 + \frac{3010}{452} = 29.5 + 6.66 \text{ or } 36.16$$

$$\sigma = \frac{i}{N} \sqrt{\Sigma fd^2x'.N - (\Sigma fdx')^2} = \frac{10}{1620} \sqrt{2306 \times 1620 - (-74)^2}$$

$$\sigma = \frac{10}{1620} \times 1931.38 = \text{Rs.}11.92$$

$$j = \frac{\bar{X} - Z}{\sigma} = \frac{34.043 - 36.16}{11.92} = \frac{-2.117}{11.92} = -0.1776$$

$$\text{C.V.} = \frac{\sigma}{X} \times 100 = \frac{11.92}{34.043} \times 100 = 35\%, \text{ So } CV = 35\% \quad J = -0.1776.$$

8.6.2 Bowley's Measure

Dr. A.L. Bowley propounded another measure of skewness based on the relative position of median and the too quartiles. If a distribution is symmetrical then Q_1 and Q_3 will be at equal distance from the Median. If a distribution is asymmetrical, the quartiles will not be equi-distant from the value of Median. Larger is the difference, higher would be the degree of skewness. Bowley's measure of skewness is called second Measure of Skewness or Quartile Measure of Skewness. This measure is useful in distributions where mode is ill-defined. Its formula is as under :

Bowley's Measure of Skewness or quartile Measure :

$$\boxed{Sk = (Q_3 - M) - (M - Q_1) \text{ or } Q_3 + Q_1 - 2M}$$

Bowley's Measure of Coefficient of Skewness or Quartile Measure :

$$\boxed{J_Q = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)} \text{ or } \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}}$$

Illustration 6 : Find out Coefficient of Skewness by Quartile Measure :

Mid-point (X)	15	20	25	30	35	40
Frequency (f)	30	28	25	24	20	21

Solution :

Calculation of Bowley's Coefficient of Skewness

M.V.	Class interval	Freq.	Cum. Freq.
15	12.5-17.5	30	30
20	17.5-22.5	28	58
25	22.5-27.5	25	83
30	27.5-32.5	24	107
35	32.5-37.5	20	127
40	37.5-42.5	21	148

$$m = \text{Size of } \frac{N}{2} \text{th item} = \text{size of } \frac{148}{2} \text{th item} = 74\text{th item}$$

Value of 74th item lies in class group 22.5-27.5

$$M = l_1 + \frac{i}{f}(m - c) = 22.5 + \frac{5}{25}(74 - 58) \text{ or } 22.5 + 3.2 = 25.7$$

$$q_1 = \text{Size of } \frac{N}{4} \text{th item} = \frac{148}{4} \text{th item} = 37\text{th item, lies in } 17.5 - 22.5 \text{ group}$$

$$Q_1 = l_1 + \frac{i}{f}(q_1 - c) = 17.5 + \frac{5}{28}(37 - 30) \text{ or } 17.5 + 1.25 = 18.75$$

$$q_3 = \text{Size of } \frac{3N}{4} \text{th item} = \frac{3(148)}{4} = 111\text{th item, lies in } 32.5 - 37.5 \text{ group.}$$

$$Q_3 = l_1 + \frac{i}{f}(q_3 - c) = 32.5 + \frac{5}{20}(111 - 107) \text{ or } 32.5 + 1 = 33.5$$

$$J_Q = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{33.5 + 18.75 - (2 \times 25.7)}{33.5 - 18.75} = \frac{52.25 - 51.4}{14.75} = 0.06$$

Illustration 7 : Find the appropriate measure of skewness and dispersion from the following data :

Age (years)	No. of Workers	Age (years)	No. of Workers
Below 20	13	35-40	112
20-25	29	40-45	94
25-30	46	45-50	45
30-25	60	50 & above	21

Solution : In open end distribution, Dr. Bowley's measure of skewness is a suitable one.

Age (years)	No. of Employees	Cum. Freq.
Below 20	13	13
20-25	29	42
25-30	46	88
30-35	60	88
35-40	112	260
40-45	94	354
45-50	45	9-399
50 & above	21	420

$q_1 = \text{Size of } \frac{420}{4}\text{th or 105th item; } m = \text{size of } \frac{420}{2}\text{th or 210th item;}$

$q_3 = \text{Size of } \frac{3(420)}{40}\text{th or 315th item. Formulae are :}$

$$Q_1 = 30 + \frac{5}{60}(105 - 88) = 30 + \frac{5 \times 17}{60} \text{ or } 31.42 \text{ years.}$$

$$M = 35 + \frac{5}{112}(210 - 148) = 35 + \frac{5 \times 62}{1120} \text{ or } 37.77 \text{ years.}$$

$$Q_3 = 40 + \frac{5}{94}(315 - 260) = 40 + \frac{5 \times 55}{44} \text{ or } 42.92 \text{ years.}$$

Bowley's Coefficient of Skewness :

$$J_Q = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{42.92 + 31.42 - 2(2 \times 37.77)}{42.92 - 31.42}$$

$$= \frac{74.34 - 75.54}{11.50} = \frac{-1.20}{11.50} = -0.104$$

Coeff. of Quartile Dispersion :

$$= \frac{Q_3 - Q_1}{Q_3 - Q_1} = \frac{42.92 - 31.42}{42.92 - 31.42} = \frac{11.5}{74.34} \text{ or } 0.154$$

So $J_Q = -0.104$ and C of Q.D. $= 0.154$

Activity C :

1. Differentiate between Bowley's measure and Karl Pearson's measure of skewness.

   _____

2. What are the limits for Bowley's co-efficient of skewness.

   _____

3. Explain which distribution is more skewed :

   Distribution I : $\overline{X} = 11,\ M = 12,\ \sigma = 5$

   Distribution II : $\overline{X} = 12.5,\ Z = 11.5,\ \sigma = 5$

   _____

8.6.3 Kelly's Measure

Kelly's measure of skewness is based on middle 90% of observatios as against Bowley's measure where middle 50% observations are taken into account. It is a mid-way between Karl Pearson's measure and Dr. Bowley's measure.

$$\text{Skewness} = P_{90} + P_{10} - 2P_{50} \text{ or } D_9 + D_1 - 2D_2$$

$$\text{Coeff. of Sk } (J_P) = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}} \text{ or } \frac{D_9 + D_1 - 2D_2}{D_9 - D_1}$$

8.6.4 Moments Measure of Skewness

This measure of skewness is based on moment about dispersion. Prof. King treated this measure as the best but its computation is very hard and complicated. It is because of lack of simplicity, that its use is restricted. It is very rarely used in actual practices. It is obtained with the help of squares, cubes etc. of the value. This measure is based on the assumption that the sum of deviations of values from its mean is zero. Its absolute measure is the cub-root of third moment of dispersion. Coefficient of skewness is obtained by

dividing third moment by standard deviation. It is also called third measure of skewness which is discussed in detail in the next unit under the head moments.

8.7 Miscellaneous Illustrations

Illustration 8 : Calculate Coefficient of Skewness in the following two distributions and tell which distribution is more skewed ?

Weight in kgs.	55-58	58-61	61-64	64-67	67-70	Total
Group A	12	17	23	18	11	81
Group B	20	22	25	13	07	87

Solution : This problem can be solved by any measure of skewness, but Karl Person's measure is more reliable and more appropriate measure, as such the same measure of skewness has been used.

Calculation of Karl Pearson's Coefficient of Skewness

Weight	A=62.5/3	Group A			Group B		
	dx'	Freq.	fdx'	fd^2x'	Freq.	fdx'	fd^2x'
55-58	-2	12	-24	48	20	-40	80
58-61	-1	17	-17	14	22	-22	22
61-64	0	23	0	0	25	0	0
64-67	1	18	18	18	13	13	13
67-70	2	11	22	44	7	14	28
Total	**-**	**81**	**-1**	**127**	**87**	**-35**	**143**

Group A : $\bar{X} = A + \frac{fdx'}{N} \times i = 62.5 \times \frac{-1}{81} \times 3 = 62.463$ kgs.

$$\sigma = \frac{i}{N}\sqrt{\Sigma fd^2x'.N - (\Sigma fdx')^2} = \frac{3}{81}\sqrt{127 \times 81 - (-1)^2}$$

$$= \frac{3}{81}\sqrt{10281 - 1} = \frac{3}{81} \times 101.39 \text{ Or } 3.76 \text{ kgs.}$$

Mode by inspection lies in $(61-64)$ group. Formula is

$$Z = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \text{ or } 61 + \frac{6}{6+5} \times 3 = 61 + \frac{18}{11} \text{ or } 62.64 \text{ kgs}$$

$$J = \frac{\bar{X} - Z}{\sigma} = \frac{62.463 - 62.64}{3.76} = -0.047$$

Group B $\bar{X} = A + \frac{fdx'}{N} \times i = 62.5 + \frac{-35}{87} \times 3$ or $62.5 - \frac{105}{87}$ or 61.293 kgs.

$$\sigma = \frac{i}{N}\sqrt{\Sigma fd^2x'.N - (\Sigma fdx')^2} = \frac{3}{87}\sqrt{143 \times 87 - (-35)^2}$$

$$= \frac{3}{87}\sqrt{12441 - 1225} \text{ or } \frac{3}{87} \times 105.906 \text{ or } 3.65 \text{ kgs.}$$

Z by inspection lies in $(61-64)$ group. Formula is

$$Z = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \text{ or } 61 + \frac{3}{3+12} \times 3 \text{ or } 61 + 0.6 \text{ or } 61.6 \text{ kgs.}$$

$$J = \frac{\bar{X} - Z}{\sigma} = \frac{61.923 - 61.6}{3.65} = -0.08$$

Hence group B is more skewed than group A.

Illustration 9 : From the following information regarding the marks obtained at College and Competitive Exams., find which group is more skewed :

College Exams.		Competitive Exams.	
Marks	**No. of Students**	**Marks**	**No. of Students**
100-150	20	1200-1250	50
150-200	45	1250-1300	85
200-250	50	1300-1350	72
250-300	25	1350-1400	60
300-350	19	1400-1450	16

Note : Use $3(\bar{X} - M)$ formula for coefficient of skewness.

Solution :

Calculation of Coefficient of Skewness of College Exams.

Marks	**Mid Values X**	***f***	**A = 225 *d′x***	***fd′x***	***fd′x²***	**c.f.**
100-150	125	20	-2	-40	80	20
150-200	175	45	-1	-45	45	65
200-250	225	50	0	0	0	115
250-300	275	25	1	25	25	140
300-350	325	19	2	38	75	159
		N=159		22	$\Sigma fd'x^2 = 226$	

Calculation of Coefficient of Skewness of Competitive Exams.

Marks	**X**	***f***	**A = 1325 *d′x***	***fd′x***	***fd′x²***	**C.F.**
1200-1250	1225	50	-2	-100	200	50
1250-1300	1275	85	-1	-85	85	135
1300-1350	1325	72	0	0	0	207
1350-1400	1375	60	1	60	60	267
1400-1450	1425	16	2	32	64	283
		N = 283		$\Sigma fd'x = 93$	$\Sigma fd'x^2 = 409$	

College Examinations

$$m = \frac{N}{2} = \frac{159}{2} = 79.5\text{th item}$$

$(200 - 250)$ group

$$M = l_1 + \frac{i}{f}(m - c)$$

$$= 200 + \frac{50}{50}(79.5 - 65)$$

$$= 200 + \frac{50 \times 145.5}{50}$$

$$= 200 + 14.5 = 214.5 \text{ marks}$$

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

$$= 225 + \frac{-22 \times 50}{159}$$

$$= 225 - 6.92 = 218.08 \text{ marks}$$

Competitive Examinations

$$m = \frac{N}{2} = \frac{283}{2} = 141.5\text{th item}$$

$(1300-1350)$ group

$$M = l_1 + \frac{i}{f}(m-c)$$

$$= 1300 + \frac{50}{72}(141.5-135)$$

$$= 1300 + \frac{50\times 6.5}{72}$$

$$= 1300 + 4.51 = 1304.51 \text{ marks}$$

$$\overline{X} = A + \frac{\Sigma fd'x}{N}\times i$$

$$= 1325 + \frac{-93\times 50}{283}$$

$$= 1325 - 16.43 = 1308.57 \text{ marks}$$

Illustration 11 : Find out coefficient of dispersion and its coefficient of skewness from the following figures. (Using Bowley's formula)

Marks	No. of Students	Marks	No. of Students
1 and not exceeding 5	10	16 and not exceeding 20	5
6 and not exceeding 10	8	21 and not exceeding 25	7
11 and not exceeding 15	12	26 and not exceeding 30	8

Solution :

Calculation of Coefficient of Dispersion & Bowley's Coefficient of Skewness

Marks	Series	*f*	*cf*
1 and not exceeding 5	0.5-5.5	10	10
6 and not exceeding 10	5.5-10.5	8	18
11 and not exceeding 15	10.5-15.5	12	30
16 and not exceeding 20	15.5-20.5	5	35
21 and not exceeding 25	20.5-25.5	7	42
26 and not exceeding 30	25.5-30.5	8	50

$$q_1 = \frac{N}{4}\text{th item} = \frac{50}{4}\text{th item} = 12.5 \text{ item, So } q_1 \text{ group is } (5.5-10.5)$$

$$Q_1 = 5.5 + \frac{5}{8}(12.5-10) = 5.5 + \frac{5}{8}\times 2.5 \text{ or } 5.5 + 1.56 = 7.06$$

$$q_3 = \frac{3N}{4}\text{th item} = \frac{3\times 50}{4}\text{th item} = 37.5\text{th item, So } q_3 \text{ group is } (20.5-25.5)$$

$$Q_3 = 20.5 + \frac{5}{7}(37.5-35) \text{ or } 20.5 + \frac{5}{7}\times 2.5 \text{ or } 20.5 + 1.786 = 22.2 \text{ marks}$$

Coeff. of Dispersion :

$$= \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} \text{ or } \frac{22.286 + 7.06 - 2\times 13.2}{22.286 - 7.06} = +0.165$$

8.8 Summary

	Measurement	Formulae
1.	Karl Pearson's Measure	$SK = \overline{X} - Z$ or $SK = 3(\overline{X} - M)$
		Coefficient of SK or J
		$J_Z = \frac{X-Z}{\sigma}$, $J_m = \frac{3(\overline{X}-M)}{\sigma}$

2. Dr. Bowley's Measure

$$SK_Q = Q_3 + Q_1 - 2M$$

$$J_Q = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

3. Kelly's Measures

$$SK_K = P_{90} + P_{10} - 2P_{50} \text{ or } D_9 + D_1 - 2D_5$$

$$J_K = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}} \text{ or } \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

8.9 Key Words

Skewness : It means lack of symmetry in the shape of a frequency distribution.

Frequency Distribution : In this the data is arranged in a systematic manner i.e. in the form of variable and frequency.

Symmetrical Distribution : In a symmetrical distribution the values of mean, median and mode coincide. The spread of frequencies is same on both sides of the centre point of the curve.

Asymmetrical Distribution : A distribution which is not symmetrical is called a skewed distribution. It is of two types positively skewed and negatively skewed.

Positively Skewed Distribution : In such a distribution the value of the mean is maximum and that of mode least - the median lies in between the two.

Negatively Skewed Distribution : In such a distribution the value of mode is maximum and that of mean least - the median lies in between the two.

8.10 Self Assessment Questions

1. Explain the meaning of skewness, comment on the various measures of skewness. Which measure is generally preferred and why ?
2. How will the skewness be tested in a distribution ? Explain with example and distinguish between dispersion and skewness.
3. Calculate the Pearson's measure of skewness on the basis of mean, mode and standard deviation:

X	:	14.5	15.5	16.5	17.5	18.5	19.5	20.5	21.5
Y	:	35	40	28	100	125	87	43	22

[Ans. $\overline{X} = 18.07$, $Z = 18.5$, $\sigma = 1.77$, $J = -0.24$]

4. Calculate Karl Pearson's second measure of coefficient of skewness from the following data :

Wages Rs.	:	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
No. of Workers	:	35	40	48	100	125	87	43	22

[Ans. $\overline{X} = 8.067$, $M = 8.5$, $\sigma = 1.776$, $J = -0.73$]

5. Calculate Karl Pearson's coefficient of skewness based on median from the following data :

Central value	:	5	10	15	20	25	30
Frequency	:	33	28	25	24	20	21

[Ans. $\overline{X} = 16.09$, $M = 15.4$, $\sigma = 8.59$, $J = +0.24$]

6. From the following data, calculate Karl Pearson's coefficient of skewness based on mode :

Weekly salary (Below Rs.) :	80	90	100	110	120	130	140	150
No. of Stenographers :	12	30	65	107	157	202	222	230

[Ans. $\overline{X} = 110.4$, Z = Rs.116.15, σ = Rs.17.26, $J = -0.33$]

7. Compute the quartile coefficient of dispersion and coefficient of skewness from the following array (use Bowley's formula) :

Central Size :	1	2	3	4	5	6	7	8	9	10
Frequency :	2	9	11	14	20	24	20	16	5	2

[Ans. $J_Q = -0.07$, Coeff. of Q.D. = 0.27]

8. Calculate the quartile coefficient of skewness from the following frequency distribution :

Weight	No. of Persons	Weight	No. of Persons
Under 100	1	150-159	65
100-109	14	160-169	31
110-119	66	170-179	12
120-129	122	180-189	5
130-139	145	190-199	2
140-149	121	200 and above	2

[Ans. $Q_1 = 124.87$, $Q_3 = 147.06$, $M = 135.71$, $J_Q = 0.023$]

9. Calculate quartiles and coefficient of skewness from the data given below :

Age under (yrs.) :	10	20	30	40	50	60
No. of Persons :	15	32	51	78	97	109

[Ans. $Q_1 = 17.21$, $Q_3 = 41.97$, $M = 31.3$, $J_Q = -0.14$]

10. Calculate measure of skewness based on quartiles and median from the following data :

Variable	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	358	2417	976	129	62	18	10

[Ans. $Q_1 = 22.625$, $Q_3 = 32.07$, $M = 26.73$, $J_Q = 0.131$]

11. Calculate quartile measure of skewness from the following :

Class	0-10	0-20	0-30	0-40	0-50	0-60	0-70	0-80	0-90	0-100
Freq.	2	5	10	20	45	65	77	87	95	100

[Ans. $Q_1 = 42$, $Q_3 = 68.33$, $M = 52.5$, $J_Q = 0.20$]

8.11 Reference Books

1. Oswal Agarwal, Bhargava, Tiwari, Meena, Business Statistics, Ramesh Book Depot, Jaipur.
2. S.P. Gupta, Statistical Methods, Sultan Chand & Sons, New Delhi.
3. Ranga, Gupta, Goyal, Bhatnager, Soni, Business Statistics, Ajmera Book Company, Jaipur.
4. Garg, Sharma, Jain & Pareek, Statistical Methods, Shivam Book House (P) Ltd. Jaipur.
5. DN Elhance, Veena Elhance & B.M. Agarwal, Fundamentals of Statistics, Kitab Mahal, Allahabad.

Unit - 9 Moments and Kurtosis

Structure of Unit:

9.0 Objectives

After completing this unit, you will be able to :

- Narrate the meaning of moments.
- Describe the objectives of studying moments
- Explain the meaning of central moments
- Calculate the value of central moments
- Calculate co-efficient based on moments
- Discuss the meaning of Kurtosis
- Explain various Measurements of Kurtosis.

9.1 Introduction

The word 'Moments' is primarily used in Physics and Mechanical Sciences. In these sciences this word is used to measure a force with respect to its tendency to produce rotation. The force of this tendency depends on two factors : The first is 'amount of force' whereas the second is 'distance from the origin at which the force is applied'. On both the sides of the fulcrum, if forces (weight × distance) are equal, there will be a balance. In case of balance, positive products (force in right side × distance of origin point to force point) and the negative products (force in left side × distance of origin point to force point) are found equal, as it depicts from the following diagram :

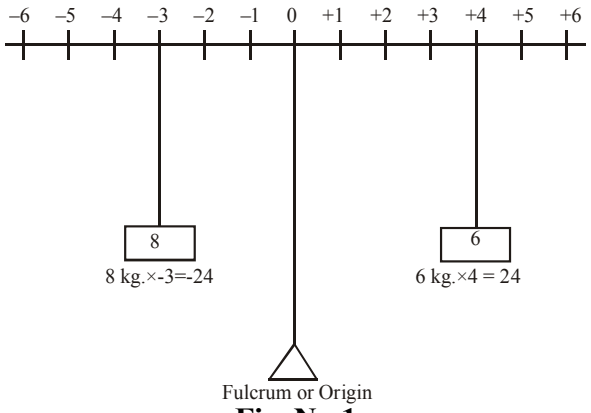


Fig. No.1

From the above diagram it is clear that if in a weighing scale, weight of 6 kg. at a distance of 4 ft. from origin point and in the other side 8 kg. weight at a distance of 3 ft. are hanged, the scale will be in a balanced position.

9.2 Meaning of Moments

In Statistics the meaning of moment is taken more or less in the same sense. Class frequencies are considered as force whereas the mean of different values or the deviations from any other point are termed here as 'distance'.

In other words, in statistics moments are the averages of 'deviations' (d^1), 'squares of deviations' `squares of deviations' (d^2), 'cubes of deviations' (d^3), and 'fourth squares of deviations' (d^4), of different values from the arithmetic mean.

If deviations are taken from actual arithmetic mean, the sum of these deviations will always be zero, if these are squared and average them we can find variance, if these are cubed their average will be skewness whereas if these are raised to fourth power and average them we find kurtosis.

9.3 Objects of Moments

The following are the main objectives of studying moments :

(1) To know about the construction of series, i.e. whether it is a symmetrical or asymmetrical frequency distribution.

(2) To study about the formation of curve of the frequency series even in case of symmetrical distribution. Normal (meso-kurtic), flat topped (play-kurtic) and peaked topped (lepto-kurtic).

9.4 Central Moments

Moments calculated from actual arithmetic mean are called as central moments or moments about the mean. The moments about mean or central moments are denoted by Greek letter μ (mue); μ_1 stands for first moment about mean, μ_2 stands for second moment about mean, μ_3 and μ_4 stand for third and fourth moments about mean, respectively. Generally in practice the first four moments are sufficient to be calculated though these can be extended to higher powers also. But these have no practical use. Yule and Kendall have rightly quoted in this respect, "Moments of higher order, though important in theory, are so extremely sensitive to sampling fluctuations that values calculated for moderate number of observations are quite unreliable and hardly ever repay the labour of computation."

9.5 Methods of Calculating Central Moments

There are two methods of calculating Central Moments :

1. Direct method
2. Short-cut method

1. Direct method :

Under this method following steps are taken :

(i) Find the arithmetic mean of the series $\overline{X}$;

(ii) Find out the deviations $(d = X - \overline{X})$;

(iii) Find out the sum of deviations, square of deviations, cubes of deviations and squares of fourth power, means $[\Sigma d, \Sigma d^2, \Sigma d^3, \Sigma d^4]$;

(iv) In case of frequency series multiply the deviations by respective frequencies and find out $\Sigma fd, \Sigma fd^2, \Sigma fd^3, \Sigma fd^4$; and ;

(v) Calculate four central moments by using the following formulae :

Individual series	Frequency series	Individual series	Frequency series
$\mu_1 = \frac{\Sigma d}{N} = \frac{\Sigma(x - \bar{X})}{N}$	$\mu_1 = \frac{\Sigma fd}{N} = 0$	$\mu_3 = \frac{\Sigma d^3}{N} = \frac{\Sigma(X - \bar{X})^3}{N}$	$\mu_3 = \frac{\Sigma d^3}{N}$
$\mu_2 = \frac{\Sigma d^2}{N} = \frac{\Sigma(X - \bar{X})^2}{N}$	$\mu_2 = \frac{\Sigma fd^2}{N} = \sigma^2$	$\mu_4 = \frac{\Sigma d^4}{N} = \frac{\Sigma(X - \bar{X})^4}{N}$	$\mu_4 = \frac{\Sigma d^4}{N}$

Illustration 1 : Compute first four moments about mean of the following data :

Student	A	B	C	D	E	F
Marks obtained	14	16	18	20	25	27

Solution :

Calculation of first moments about the mean

Students	Marks obtained	$(X - \bar{X})$ d	d^2	d^3	d^4
A	14	– 6	36	– 216	1,296
B	16	– 4	16	– 64	256
C	18	– 2	4	– 8	16
D	20	0	0	0	0
E	25	+ 5	25	+ 125	625
F	27	+ 7	49	+ 343	2401
Total	120	0	130	+ 180	4,594

$$\mu_1 = \frac{\Sigma d}{N} = \frac{0}{6} = 0; \ \mu_3 = \frac{\Sigma d^3}{N} = \frac{180}{6} = 30 \quad \bar{X} = \frac{\Sigma X}{N} = \frac{120}{6} = 20$$

$$\mu_2 = \frac{\Sigma d^2}{N} = \frac{130}{6} = 21.67; \ \mu_4 = \frac{\Sigma d^4}{N} = \frac{4594}{6} = 765.6$$

Illustration 2 : Calculate first four moments about the mean by Direct Method from the following data :

X :	10	12	14	16	18
F :	4	4	2	8	2

Solution :

(X)	(f)	(fX)	$d = (x - \bar{x})$	$f.d^1$	$f.d^2$	$f.d^3$	$f.d^4$
10	4	40	– 4	–16	64	– 256	1024
12	4	48	– 2	– 08	16	– 32	64
14	2	28	0	0	0	0	0
16	8	128	+ 2	+ 16	32	+ 64	128
18	2	36	+ 4	+ 8	32	+ 128	512
Total	20	280		0	144	-96	1728

$$\bar{X}(\text{mean}) = \frac{\Sigma(fX)}{N} = \frac{280}{20} \text{ or } \bar{X} = 14$$

$$\mu_1 = \frac{\Sigma fd^1}{N} = \frac{0}{20} = 0; \ \mu_2 = \frac{\Sigma fd^2}{N} = \frac{144}{20} = 7.2$$

$$\mu_1 = \frac{\Sigma fd^1}{N} = \frac{-96}{20} = -4.8; \quad \mu_4 = \frac{\Sigma fd^4}{N} = \frac{1728}{20} = 86.4;$$

Illustration 3 : Calculate first four moments about the mean by Direct Method from the following data :

(x) :	0 – 10	10 – 20	20 – 30	30 – 40
(f) :	2	4	6	8

Solution : **Calculation of first four moments (Direct Method)**

(x)	*(f)*	*(x)*	*(fx)*	$\bar{x} = 25$ $d = x - \bar{x}$	$f.d^1$	$f.d^2$	$f.d^3$	$f.d^4$
0 – 10	2	5	10	– 20	– 40	800	– 16,000	3,20,000
10 – 20	4	15	60	– 10	– 40	400	– 4,000	40,000
20-30	6	25	150	0	0	0	0	0
30-40	8	35	280	+ 10	+ 10	800	+ 8,000	80,000
Total	20		500		0	2,000	– 12000	4,40,000

$$\text{Mean } (\bar{X}) = \frac{\Sigma fx}{N} \text{ or } \frac{500}{200} = 25$$

$$\mu_1 = \frac{\Sigma fd}{N} \text{ or } \frac{0}{20} = 0; \qquad \mu_3 = \frac{\Sigma fd^3}{N} \text{ or } \frac{12,000}{20} = 600$$

$$\mu_2 = \frac{\Sigma fd^2}{N} \text{ or } \frac{2000}{20} = 100; \qquad \mu_4 = \frac{\Sigma fd^4}{N} \text{ or } \frac{4,40,000}{20} = 22,000$$

2. Short-cut method :

If the arithmetic mean is not in whole number but it is in fraction, in this situation it is better to use the short-cut method to avoid mathematical complications. In this method—

(i) Assume an arbitrary mean 'A'.

(ii) Find the deviations (dx) of the values from assumed mean, and also their squares (dx^2), cubes (dx^3) and fourth square (dx^4).

(iii) In case of frequency series the above deviations and their squares etc. are multiplied by respective frequencies and find out Σfdx, Σfdx^2 , Σfdx^3 and Σfdx^4.

(iv) Thereafter, the following formulae are used for computing first four moments. These moments are called as 'moments about arbitrary origin'. These are denoted by Greek letter ν (NUE) : $\nu_1, \nu_2 \nu_3, \nu_4$.

Individual Series	**Frequency Series**
$v_1 = \frac{\Sigma dx}{N} = \frac{\Sigma(X-A)}{N}$	$v_1 = \frac{\Sigma fdx}{N} = \frac{\Sigma f(X-A)}{N}$
$v_2 = \frac{\Sigma d^2x}{N} = \frac{\Sigma(X-A)^2}{N}$	$v_2 = \frac{\Sigma fdx^2}{N} = \frac{\Sigma f(X-A)^2}{N}$
$v_3 = \frac{\Sigma d^3x}{N} = \frac{\Sigma(X-A)^3}{N}$	$v_3 = \frac{\Sigma fd^3x}{N} = \frac{\Sigma f(X-A)^3}{N}$
$v_4 = \frac{\Sigma d^4x}{N} = \frac{\Sigma(X-A)^4}{N}$	$v_4 = \frac{\Sigma fdx^4}{N} = \frac{\Sigma f(X-A)^4}{N}$

(v) In the end by using the following formulae 'moments about the mean' are calculated from the 'moments about the arbitrary origin' :

$$\mu_1 = \nu_1 - \nu_1 = 0; \qquad \mu_2 = \nu_2 - \nu_1^2 = \sigma^2$$

$$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3; \quad \mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2, \nu_1^2 - 3\nu_1^4$$

Illustration 4 : Calculate first four moments about the mean by short-cut method from the following data :

Length (in inches)	1.0	2.0	3.0	4.0	5.0	6.0	7.0
Frequency	5	38	65	92	70	40	10

Solution : **Calculation of First Four Moments**
(Short-cut method)

Length in inches	Frequency	dx (A = 4.0)	fdx	fdx^2	fdx^2	fdx^4
1.0	5	– 3	– 15	45	– 135	405
2.0	38	– 2	– 76	152	– 304	608
3.0	65	– 1	– 65	65	– 65	65
4.0	92	0	0	0	0	0
5.0	70	+1	+ 70	70	+ 70	70
6.0	40	+2	+ 80	160	+ 320	640
7.0	10	+3	+ 30	90	+270	810
Total	320	–	+ 24	+ 582	+ 156	+ 2598

Moments about an arbitrary origin :

$$\nu_1 = \frac{\Sigma fdx}{N} = \frac{24}{320} = +0.075; \qquad \nu_3 = \frac{\Sigma fd^3x}{N} = \frac{156}{320} = +0.488$$

$$\nu_2 = \frac{\Sigma fd^2x}{N} = \frac{582}{320} = 1.819; \qquad \nu_4 = \frac{\Sigma fd^4x}{N} = \frac{2598}{320} = 8.119$$

Moments about the Mean :

$$\mu_1 = \nu_1 - \nu_1 \text{ or } .075 - .075 = 0$$

$$\mu_2 = \nu_2 = \nu_1^2 \text{ or } 1.819 - (.075)^2 = 1.813$$

$$\mu_3 = \nu_3 - 3\nu_2, \nu_1 + 2\nu_1^3 \text{ or } 0.488 - 3(1.819 \times 0.075) + 2(.075)^3$$

$$= 0.488 - 0.409 + 0.00084 = .0798$$

$$\mu_4 = \nu_4 - 4\nu_3.\nu_1 + 6\nu_2.\nu_1^2 - 3\nu_1^4$$

$$= 8.119 - 4.(.488 \times .075) \times 6(1.819)(0.075)^2 - 3(.075)^4$$

$$= 8.119 - 0.146 + 0.0614 - 0.00060 \text{ or } 8.034.$$

Activity A :

1. Explain the third and fourth central moments in terms of the first four moments about the origin.

   _____

2. In a symmetrical distribution which moment would always be zero.

   _____

3. How moments help in analyzing a frequency distribution.

   _____

9.6 Computation of Moments about Arbitrary Origin and From Central Moments

On the basis of central moments the moments about any arbitrary origin also called 'raw moments' can be computed by using following steps :

(i) Find the difference between arithmetic mean $(\overline{X})$ and assumed mean (A) :

$$\left[\overline{d}x = (\overline{X} - A)\right]$$

(ii) Thereafter the following formulae are used :

$$\nu_1 = (\mu + \bar{d}x)^1 = \mu_1 + \bar{d}_x \ (\because \mu_1 = 0) \qquad \text{so} \qquad \nu_1 = \bar{d}x$$

$$\nu_2 = (\mu + \bar{d}x)^2 = \mu_2 + 2\mu_1\bar{d}_x + \bar{d}x^2 = \mu_2 + \bar{d}x^2 (\because \mu_1 = 0)$$

$$\nu_3 = (\mu + \bar{d}x)^3 = \mu_3 + 3\mu_2\bar{d}x + 3\mu_1\bar{d}x^2 + \bar{d}x^3 \text{ or } = \mu_3 + 3\mu_2\bar{d}x + \bar{d}x^3$$

$$\nu_4 = (\mu + \bar{d}x)^4 = \mu_4 + 4\mu_2\bar{d}x + 6\mu_2\bar{d}x^2 + 4\mu_1\bar{d}x^3 + \bar{d}x^4$$

or $\quad \mu_4 + 4\mu_3\bar{d}x + 6\mu_2\bar{d}x^2 + \bar{d}x^4 \ (\because \mu_1 = 0)$

Note : $(\bar{X} - A)$ has been denoted here as $\bar{d}x$, it may be denoted by (Greek sign) Δ (delta) also.

Illustration 5 : From the following first four moments of a distribution about the arbitrary origin 4, find out the mean of the distribution and calculate moments about the mean and also about the arbitrary mean zero.

$$\nu_1 = 1, \nu_2 = 4, \nu_3 = 10 \text{ and } \nu_4 = 45 .$$

Solution : $\quad \bar{X} = A + \dfrac{\Sigma dx}{N}$ and $\nu_1 = \dfrac{\Sigma dx}{N}$ So, $\bar{X} = A + \nu_1$

In the problem A = 4 and $\nu_1 = 1$ Hence $\bar{X} = 4 + 1 = 5$

(i) First Four Moments about the Mean $(\bar{X} = 5)$

$$\mu_1 = \nu_1 - \nu_1 = 1 - 1 = 0$$

$$\mu_2 = \nu_2 - \nu_1^2 = 4 - (1)^2 = 4 - 1 = 3$$

$$\mu_4 = \nu_4 - 4\nu_3.\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4 \text{ or } 45 - 4(10 \times 1) + 6(4)(1)^2 - 3(1)^4$$

$$= 45 - 40 + 24 - 3 = 26, \qquad \text{So } \bar{X} = 5, \mu_1 = 0, \mu_2 = 3,, \ \mu_3 = 0, \ \mu_4 = 26$$

(ii) First Four Moments about Arbitrary Origin Zero (0) :

Since $\quad \bar{X} = 5, A = 0, \text{ So } \bar{d}x = (\bar{X} - A) = (5 - 0) = 5$

$$\nu_1 = \mu_1 + \bar{d}x = 0 + 5 = 5; \nu_2 = \mu_2 + \bar{d}x^2 = 3 + (5)^2 = 28$$

$$\nu_3 = \mu_3 + 3\mu_2\bar{d}x + \bar{d}x^3 = 0 + 3(3)(5) + (5)^2 \quad \text{or} \quad 0 + 45 + 125 = 170$$

$$\nu_4 = \mu_4 + 4\mu_3\bar{d}x + 6\mu_2\bar{d}x^2 + \bar{d}_x^4 = 26 + 4(0)(5) + 6(3)(5)^2 + (5)^4$$

$$= 26 + 0 + 450 + 625 = 1101$$

So, $\nu_1 = 5, \nu_2 = 28, \nu_3 = 170$ and $\nu_4 = 1101$.

9.7 Sheppard's Correction

In case of continuous series, while computing moments, it is assumed that all the frequencies are concentrated at the centre or middle point of the class groups. But this assumption is not found correct and it introduces some error which is known as grouping error. For eliminating these errors Sheppard's (correction) method is used by applying the following formulae :

$$\mu_1(\text{Corrected}) = \mu_1(\text{Correction is not required})$$

$$\mu_2(\text{Corrected}) = \mu_2(\text{Uncorrected}) - \frac{i^2}{12}$$

$$\mu_3(\text{Corrected}) = \mu_2(\text{Correction is not required})$$

$$\mu_4(\text{Corrected}) = \mu_4(\text{Uncorrected}) - \frac{1}{2}\mu_2(\text{Uncorrected})\, i^2 + \frac{7}{240} i^4$$

Conditions for Sheppard's Correction :

(i) This correction is to be applied in continuous series (continuous frequency distribution) only.

(ii) Sheppard's correction is not required in case of first and third moments because positive and negative signs of deviations remain in them. Thus in this situation the error being in compensatory nature, which is automatically removed.

(iii) In case of second and fourth moments deviations are squared and raised to fourth power hence the negative signs become positive signs, which help in committing the error, so Sheppard's correction is required.

(iv) When the frequency distribution is significantly skewed i.e. J-shaped or U-shaped, correction is not required.

Activity B

1. What is the purpose of Sheppard's correction.

   _____

2. Sheppard's correlation is not applicable to which type of distribution.

   _____

3. You are given the following values of moments :

   $\mu_2 = 43.553,\ \mu_3 = -9.774$ *and* $\mu_4 = 5508.567$

   Find the corrected values of each one of these, taking into account the class interval which is 3.

Illustration 6 : Compute first four moments of the following data and make Sheppard's correction also, if necessary :

Marks (less than)	20	30	40	50	60	70	80
No. of Students	1	21	90	198	276	298	300

Solution : Cumulative frequencies are given in the question. So for solving the question simple frequencies will be computed.

Calculation of first four moments about mean

Marks	No. of Students	Mid-Values	d′x, A = 45	fd′x	fd′x^2	fd′x^3	fd′x^4
10-20	1	15	-3	-3	9	-27	81
20-30	20	25	-2	-40	80	-160	320
30-40	69	35	-1	-69	69	-69	69
40-50	108	45	0	0	0	0	0
50-60	78	55	1	78	78	78	78
60-70	22	65	2	44	88	176	352
70-80	2	75	3	6	18	54	162
	N = 300			+ 6	342	+ 52	1,062

$$v_1' = \frac{\Sigma fd'x}{N} = \frac{16}{300} \text{ or } 0.053;$$

$$v_2' = \frac{\Sigma fd'x^2}{N} = \frac{52}{300} \text{ or } 0.173;$$

$$v_3' = \frac{\Sigma fd'x^3}{N} = \frac{342}{300} \text{ or } 1.14$$

$$v_4' = \frac{\Sigma fd'x^4}{N} = \frac{1062}{300} \text{ or } 3.54$$

Movements about Mean (Central Moments) :

$$\mu_1 = \left(v_1' - v_1'\right) \times i = (.053 - .053) \times 10 = 0$$

$$\mu_2 = \left(v_2' - v_2'^2\right) \times i^2 = (1.14 - .002809) \times 100 = 113.72$$

$$\mu_3 = \left(v_3' - 3v_2'v_1' + 2v_1^3\right) \times i^3$$

$$= \left[(.172) - (3 \times 1.14 \times .053) + 2(.053)^3\right] \times 10^3$$

$$= (.172 - .18126 + .000297754) \times 1000 \text{ or } -8.962$$

$$\mu_4 = \left(v_4' - 4v_3'\, v_1' + 6v_2'\, v_1'^2 - 3v_1'^4\right) \times i^4$$

$$= \left[3.54 - (4 \times .173 \times .053) + \left(6 \times 1.14 \times .053^2\right) - 3(.053)^4\right] \times (10)^4$$

$$= [3.54 - .36676 + .01921356 - .00002367] \times 10000$$

or $= 35224$

Sheppard's Correction : μ_1 and μ_3 **are not required to be corrected.**

$$\mu_2 \text{ (Corrected)} = \mu_2 - \frac{i^2}{12} = 113.72 - \frac{10^2}{12} = 105.39$$

$$\mu_4 \text{ (Corrected)} = \mu_4 - \frac{\mu_2 \times i^2}{2} + \frac{7i^4}{240}$$

$$= 31924.30 - \frac{(113.72 \times 100)}{2} + \frac{7 \times 10000}{240} \text{ or } 31924.30 - 5686 + 291.67 = 26,529.97$$

Charlier's Check of Accuracy :

Charlier's check of accuracy formulae can be used for testing the accuracy of the computation of moments

First Moments	:	$\Sigma f(dx + 1) = \Sigma fdx + N$
Second Moments	:	$\Sigma f(dx + 1)^2 = \Sigma fd^2x + 2\Sigma fdx + N$
Third Moments	:	$\Sigma f(dx + 1)^3 = \Sigma fd^3x + 3\Sigma fd^2x + 3\Sigma fdx + N$
Fourth Moment	:	$\Sigma f(dx + 1)^4 = \Sigma fd^4x + 4\Sigma fd^3x + 6\Sigma fd^2x + 4\Sigma fdx + N$

9.8 Coefficient Based on Moments

On the basis of the relationship between different moments Alpha coefficient (α), Beta coefficient (β) and Gamma coefficient (γ) can be computed by using following formulae :

Alpha Coefficient	Beta Coefficient	Gamma Coefficient
$\alpha_1 = \frac{\mu_1}{\sigma} = 0$	$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \alpha_3^3$	$\gamma_1 = \sqrt{\beta_1} = \alpha_3$
$\alpha_1 = \frac{\mu_2}{\sigma^2} = 1$	$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \alpha_3$	$\gamma = \beta_2 - 3 = \frac{\mu_4 - 3\mu_2^2}{\mu_2^2}$
$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}$	$\beta_2 = \frac{\mu_4}{\mu_2^2} = \alpha_4$	—
$\alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$	—	—

From the above Coefficients some important moments are presented in the form of Moment Ratios and are used in the measurement of skewness and kurtosis.

9.9 Measures of Skewness Based on Moments

The odd moments about mean in a symmetrical distribution are $\mu_1, \mu_3, \mu_5, \mu_7$, their values will always be zero.

According to Karl Pearson there are following two formulae for computation of coefficient of skewness which is based on moment ratios.

First Coefficient of Skewness

$$\sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} \text{ or } \frac{\mu_3}{\mu_2^{3/2}} \text{ or } \alpha_3 \text{ or } \gamma_1$$

Here, positive and negative signs are not ignored. This measurement of skewness is called as First Coefficient of Skewness.

Second Coefficient of Skewness :

When the skewness in a series exists in a very low degree, it is better to compute second coefficient of skewness. Its formula is :

Second Coefficient of Skewness

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

With the help of this formula 'Mode' can also be computed :

$$Z = \overline{X} - \left[\sigma \times \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}\right] \text{ Since } \frac{\overline{X} - Z}{\sigma} = \frac{\beta_1(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

or $$\overline{X} - Z = \sigma \times \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \text{ or } Z = \overline{X} - \left[\sigma \times \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}\right]$$

Illustration 7 : Calculation Coefficient of Skewness based on third moment for the followig distribution :

Weight (kg.)	80-85	85-90	90-95	95-100	100-105	105-110	110-115	115-120	120-125
No. of persons	7	31	42	54	33	24	22	8	4

Solution :

Calculation of Coefficient of Skewness

Weight (kg)	Mid-value	No. of Persons	d′x	fd′x	fd′x^2	fd′x^3
80-85	82.5	7	-3	-21	63	-189
85-90	87.5	31	-2	-62	124	-248
90-95	92.5	42	-1	-42	42	-42
95-100	97.5	54	0	0	0	0
100-105	102.5	33	+1	33	33	+33
105-110	107.5	24	+2	48	96	+192
110-115	112.5	22	+3	66	198	+594
115-120	117.5	8	+4	32	128	+512
120-125	112.5	4	+5	20	100	+500
		N = 225		+74	+784	+1352

$$v_1' = \frac{\Sigma fd'x}{N} = \frac{74}{225} = 0.3288; \qquad v_2' = \frac{\Sigma fd'x^2}{N} = \frac{784}{225} = 3.4844$$

$$v_3' = \frac{\Sigma fd'x^3}{N} = \frac{1352}{225} = 6.0088;$$

$$\mu_1 = \left(v_1' - v_1'\right) \times i \text{ or } (.3288 - .3288) \times 5 = 0$$

$$\mu_2 = \left(v_2' - v_1'^2\right) \times i^2 = \left(3.4844 - 0.3288^2\right) \times 5^2$$

$$= (3.4844 - 0.1081) \times 25 \text{ or } 84.41$$

$$\mu_3 = \left(v_3' - 3v_1'\, v_2 + 2v_1'^3\right) \times i^3$$

$$= \left[6.0018 - (3 \times 0.3288 \times 3.4844) + 2 \times .3288^3\right] \times 5^3$$

$$= [6.0018 - 3.4370 + .0711] \times 125 \text{ or } 2.6358 \times 125 = 329.4875$$

Coefficient of skewness based on moments $(\beta_1) = \frac{\mu_4^2}{\mu_2^3}$

$$= \frac{(329.49)^2}{(84.41)^3} \text{ or } \frac{108563.66}{601425.47} = 0.18$$

9.10 Kurtosis

Kurtosis is a statistical measure which tells about the degree of flatness or peakedness in the region of the mode of a frequency curve. In Greek language the word kurtosis refers to "bulginess".

According to **Simpson** and **Kafka**, "The degree of Kurtosis of a distribution is measured relative to the peakedness of a normal curve."

Croxton and **Cowden** say, "A measure of Kurtosis indicates the degree to which a curve of the frequency distribution is peaked or flat topped."

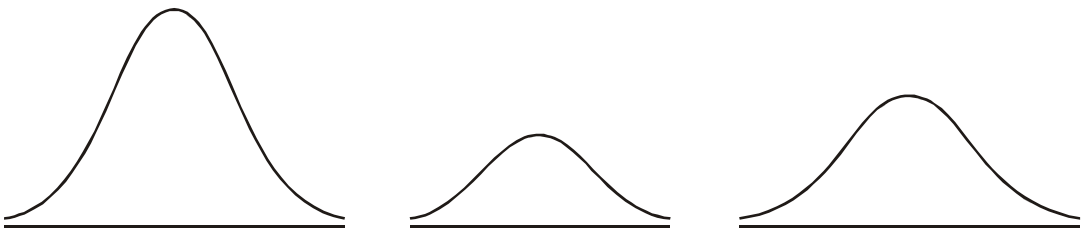
In the words of **Clark** and **Shakade**, "Kurtosis is the property of a distribution which expresses relative peakedness."

The following three words were used by Karl Pearson in 1905 :

1. LEPTOKURTIC : Peaked curve
2. PLATYKURTIC : Flat-topped curve
3. MESOKURTIC : Normal Curve

Leptokurtic : If a curve is more peaked than the normal curve, it is called 'leptokurtic'. In such a case the values are more concentrated near the mode, hence the shape of this curve finds like Kangaroo **[see Fig. a].**

Platykurtic : If a curve is more flat-topped than the normal curve it is called as 'platykurtic' [see **Fig. b].**



(a) Leptokurtic *(b) Platy kurtic* *(c) Mesokurtic*

Mesokurtic : The normal curve itself is termed as 'mesokurtic', its shape is like a bell. [see **Fig. c].**

9.11 Measurement of Kurtosis

The measurement of kurtosis is based on fourth and second moment. According to Karl Pearson's formula:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

(i) If $\beta_2 = 3,$ the curve is said to be normal or Meso Kurtic.

(ii) If $\beta_2 > 3,$ the curve is peak-topped and termed as Leptokurtic.

(iii) If $\beta_2 < 3,$ the curve is flat-topped and termed as Platykurtic.

γ_2 (Gamma two) may also be used for measurement of Kurtosis—

If γ_2 i.e. $\beta_2 - 3 = 0,$ curve is normal or Mesokurtic.

If γ_2 is positive, the curve is Leptokurtic.

If γ_2 is negative, the curve is Platykurtic.

Activity C

1. With the help of diagrams show the various types of kurtosis, and also indicate their name.
2. If $\beta_2 = 3,$ the distribution is called .....

   If $\beta_2 > 3,$ the distribution is .....

   If $\beta_2 < 3,$ the distribution is .....
3. Point out the role of kurtosis in analysing a frequency distribution.
4. The following data are given to an economist for the purpose of economic analysis. The data refer to the length of life of a sample of Good Year Tyres.
   Is the distribution platykurtic ?

   $N = 100,\ \Sigma fdx = 50,\ \Sigma fdx^2 = 1967.2$

   $\Sigma fdx^3 = 2925.8$ *and* $\Sigma fdx^4 = 86,650.2$

Illustration 8 : First four central moment of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and kurtosis of the distribution.

Solution :

For Skewness : $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = \frac{0.49}{15.625} = +0.03$

For Kurtosis : $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2}$ or 3

As $\beta_2 = +0.03,$ which indicates that the distribution is skewed but very insignificant. In the same way $\beta_2 = 3$, which indicates that the distribution is normal or Mesokurtic.

Illustration 9 : The four moments of frequency distribution about an arbitrary origin are : $v_1 = -2,\ v_2 = 14,\ v_3 = 20,$ and $v_4 = 50.$ Find out the value of β_1 and β_2.

Solution : The value of β_1 and β_2 are based on central moments, hence at first central moment shall be computed from the moments about arbitrary origin :

$$\mu_1 = v_1 - v_2, = (-2) - (-2) = 0$$

$$\mu_2 = v_2 - v_1, = 14 - (-2)^2 \text{ or } 14 - 4 = 10$$

$$\mu_3 = v_3 - 3v_2v_1, + 2v_1^3 \text{ or } -20 - (3 \times 14 \times -2) + 2(-2)^3 \text{ or } -20 + 84 - 16 \text{ or } 48$$

$$\mu_4 = v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4$$

$$= 50 - (4 \times -20 \times -2) + \left[6 \times 14 \times (-2)^2\right] - 3(-2)^4$$

or $\quad 50-160+336-48=178$

$$\beta_1\left(\text{Skewness}\right) = \frac{\mu_3^2}{\mu_2^3}=\frac{(48)^2}{(10)^3} \text{ or } \frac{2304}{1000} = 2.304$$

$$\beta_2\left(\text{Kurtosis}\right) = \frac{\mu_4}{\mu_2^2}=\frac{178}{(10)^2} \text{ or } \frac{178}{100} = 1.78$$

It is evident from the above that the distribution is skewed and not symmetrical as the value of β_2 is 2.304. In the same way the distribution is Platykurtic because the value of β_2 is less the an three $(\beta_2 < 3)$.

9.12 Summary

Direct Method (deviation from actual mean)

$$\mu_1=\frac{\Sigma d}{N}=\frac{\Sigma(X-\overline{X})}{N} \qquad \mu_1=\frac{\Sigma fd}{N}=0$$

$$\mu_2=\frac{\Sigma d^2}{N}=\frac{\Sigma(X-\overline{X})^2}{N} \qquad \mu_2=\frac{\Sigma fd^2}{N}=\sigma^2 \text{ variance}$$

$$\mu_3=\frac{\Sigma d^3}{N}=\frac{\Sigma(X-\overline{X})^3}{N} \qquad \mu_3=\frac{\Sigma fd^3}{N}$$

$$\mu_4=\frac{\Sigma d^4}{N}=\frac{\Sigma(X-\overline{X})^4}{N} \qquad \mu_4=\frac{\Sigma fd^4}{N}$$

Short-cut Method (deviations from assumed mean)

$$V_1=\frac{\Sigma dx}{N}=\frac{\Sigma(X-A)}{N} \qquad V_1=\frac{\Sigma fdx}{N}=\frac{\Sigma f(X-A)}{N}$$

$$V_2=\frac{\Sigma d^2x}{N}=\frac{\Sigma(X-A)^2}{N} \qquad V_2=\frac{\Sigma fd^2x}{N}=\frac{\Sigma f(\overline{X}-A)^2}{N}$$

$$V_3=\frac{\Sigma d^3x}{N}=\frac{\Sigma(X-A)^3}{N} \qquad V_3=\frac{\Sigma fd^3x}{N}=\frac{\Sigma f(\overline{X}-A)^3}{N}$$

$$V_4=\frac{\Sigma fd^4x}{N}=\frac{\Sigma(X-A)^4}{N} \qquad V_4=\frac{\Sigma fd^4x}{N}=\frac{\Sigma f(\overline{X}-A)^4}{N}$$

Step deviation method (for equal class intervals)

$$\mu_1=\frac{\Sigma fd'x}{N}\times i \qquad \mu_2=\frac{\Sigma fd'x^2}{N}\times i^2 \qquad \mu_3=\frac{\Sigma fd'x^3}{N}\times i^3 \qquad \mu_4=\frac{\Sigma fd'x^4}{N}\times i^4$$

Applying correction to find out moments about mean or central moments :

$$\mu_1=v_1-v_2=0 \qquad \mu_2=v_2-v_1^2=\sigma^2$$

$$\mu_3=\nu_3-3\nu_2\nu_1+2\nu_1^3 \qquad \mu_4=\nu_4-4\nu_3\nu_1+6\nu_2v_1^2-3\nu_1^4$$

Shepherd's correction (in continuous series) : μ_1 and μ_3 (No correction needed)

Corrected $\mu_2=\mu_2-\frac{i^2}{12}$ (i = class interval); Corrected $\mu_4=\mu_4-\left[\frac{\mu_2\times i^2}{2}+\frac{7}{240}i^4\right]$

Skewness : $\beta_1=\frac{\mu^3}{\mu_2^3}$ or $\sqrt{\beta_1}=\frac{\mu_3}{\mu_3^{1/2}}=\frac{\mu^3}{\sigma^3}$, Kurtosis $\beta_2=\frac{\mu_4}{\mu_2^2}=\frac{\mu^4}{\sigma^4}$.

If $\beta_2=3$ or $\gamma_2=\beta_2-3=0$ (MesoKurtic) If $\beta_2>3$ or γ_2 is positive (Lepto-Kurtic). If $\beta_2=<3$ or γ_2 is negative (Platy Kurtic).

9.13 Self Assessment Questions

1. Define moments in statistics. Describe the various methods of calculating moments.
2. What is kurtosis? What purpose does it serve ? Is the study of kurtosis useful in economics and social sciences ? If not, why?
3. Differentiate between Moments and Kurtosis.
4. What is Sheppard's correction? In which series is it required.
5. State the various types of Kurtosis.
6. The first four moments from the value 2 of a distribution are 1, 2.5, 5.5 and 16. Calculate first four central moments.

   Ans. $\mu_1 = 0,\ \mu_2 = 1.5,\ \mu_3 = 0,\ \mu_4 = 6$

7 The first moment fo distribution about the value 4 (A = 4) are respectively -1.5, 17, -30 and 108. Find out moments about the mean and about the origin 0 (zero).

   Ans. About; About (0) : 2.5, 21, 166, 1132

8. First four central moments of a distribution are 0, 3, 0 and 26. Calculate first four moments : (i) about arbitrary origin 4, and (ii) about zero (0).

   Ans. About (4) : 1, 4, 10, 45; About (zero) : 5, 28, 170, 1101.

9. The first four moments of a distribution are 1, 4, 10 and 46 respectively. Calculate the first four central moments and the beta coefficients. Comment on the nature of the distribution.

   Ans. 0, 3, 0, 27; $\beta_1 = 0,\ \beta_2 = 3.$ Perfectly symmetrical and Mesokurtic.

10. (i) The second, third and fourth moments of a variety are 19.67, 29.26 and 866 respectively. Find out beta-coefficients.

    (ii) Find the corrected moments by applying Sheppard's correction of the following value if the magnitude of the class interval is 3.

    Given : $\mu_2 = 43.535,\ \mu_3 = -9.774,\ \mu_4 = 5508.567;$

    Ans. (i) $\beta_1 = .1126,\ \beta_2 = 2.239;$ (ii) Corrected $\mu_2 = 42.785,\ \mu_4 = 5310.297.$

11. Calculate first four moments about the mean from the following data and apply Sheppard's correction if necessary :

Value	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	1	20	69	108	78	22	2

    Ans. $\mu_1 = 0,\ \mu_2 = 113.72,\ \mu_3 = -8.69,\ \mu_4 = 31924.30;$ Correction $\mu_2 = 105.4,\ \mu_4 = 26529.47.$

12. From the following age-statistics fo 250 persons in a sample study, find the coefficient of skewness with the help of second and third moments about the mean.

Age (Less than)	10	20	30	40	50	60	70	80	90
No. of Persons	15	35	60	84	96	127	188	200	250

    Note : Assume 45 as arbirary mean (A)

    Ans. $\mu_3 = -5939.62,\ \mu_2 = 629.96,\ \sqrt{\beta_1} = -.376$

9.14 Reference Books

1. Oswal, Agarwal, Bhargava, Tiwari Meena, Business Statistics, Ramesh Book Depot, Jaipur.
2. S.P. Gupta, Statistical Methods Sultan Chand & Sons, New Delhi.
3. Ranga, Gupta, Goyal, Bhatnager, Soni, Business Statistics, Ajmera Book Company, Jaipur.
4. Garg, Sharma, Jain & Pareek, Statistical Methods, Shivam Book House (P) Ltd. Jaipur.
5. DN Elhance, Veena Elhance & B.M. Agarwal, Fundamentals of Statistics, Kitab Mahal, Allahabad.

Unit - 10 Index Number

Structure of Unit:

10.0 Objectives

After completing this unit, you will be able to:

- Understand that index numbers describe how much economic variables have changed over time;
- Point out various uses of index numbers;
- Know and avoid problems in constructing index numbers;
- Become familiar with the three principal types of index: price index, quantity index, and value index;
- Learn how to calculate various kinds of index numbers;
- Describe various limitations of index numbers.

10.1 Introduction

In our day to day life, things keep changing. The prices of various commodities vary at some rate over a period of time. In order to be update on such price changes, we need some methods to predict these. In business area for the budget purpose a manager may be interested to know how the raw material prices have increased over last one year. May be some changes in price; indicate the trend of increase on the regular basis, so that future planning could be more accurate. For such variations, we may analyze the degree of change in the form of Index Numbers.

Index numbers are indicators which reflect the relative changes in the level of a certain phenomenon in any given period (or over a specified period of time) called the 'current period' with respect to its values in some fixed period, called the 'base period' selected for comparison.

Index numbers are often known as the barometers of economic activity as they help to get an idea of the present day situation with regard to changes in production, consumption, exports and imports, national income, business level, cost of living, the price of a particular commodity or a group of commodities, Industrial or agricultural production, stocks and shares, sales and profits of a business house, volume of trade, factory production, wage structure of workers in various sectors, bank deposits, foreign exchange reserves, and so on.

Some of the important definitions of index numbers are given as under:

According to Wessell, Willet and Simone, "An index number is a special type of average that provides a measurement of relative changes from time to time or from place to place".

According to Edgeworth, "Index number shows by its variations the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice".

According to Murray R. Spiegel, "An Index Number is a statistical measure designed to show changes in variable or a group of related variables with respect to time, geographic location or other characteristics."

According to Croxton and Lowden, "Index Numbers are devices for measuring differences in the magnitude of a group of related variables."

According to Karmal and Pollasek, "An index number is a device for comparing the general level of magnitude of a group of distinct, but related, variables in two or more situations".

According to Patterson, "In its simplest form, an index number is the ratio of two index numbers expressed as a percent. An index number is a statistical measure - a measure designed to show changes in one variable or in group of related variable over time or with respect of geographic location or other characteristics".

According to Dr. A. L. Bowley, "A series of index numbers is a series which reflects in its trend and fluctuations, the movements of some quantity to which it is related."

According to John I. Griffin, "An Index Number is a quantity which by reference to a base period shows by its variations the changes in the magnitude over a period of time. In general, Index Numbers are used to measure changes over time in magnitudes which are not capable of direct measurement."

Thus, it is apparent from above definitions that an index number is a statistical device which measures the extent to which a group of related variable changes over a period of time. Index number in fact relates a variable or group of variables in a given period to the same group of variables in some other period.

Some of the important characteristics of index numbers include the following:

- Index numbers are the specialized averages.
- Index numbers record the net changes in a group of related variables over a period of time.
- Index numbers measure changes not capable of direct measurement.
- Index numbers are for comparison.
- Index numbers are expressed in percentages.

Activity A:

1. "An Index number is a special type of average". Discuss.
2. According to you which definition of index number is more appropriate and why?

10.2 Uses of Index Numbers

The first index number was constructed by an Italian, Mr. Carli, in 1764 to compare the changes in price for the year 1750 (current year) with the price level in 1500 (base year) in order to study the effect of price level in Italy. Though originally designed to study the general level of prices or accordingly purchasing power of money, today index numbers are an important tool of economic analysis and they reveal the pulse of the economy.

Use of Index numbers is the most powerful tool in the hands of management, government officials and individuals to analyse the business and economic situations of a country. Some of the important uses or significance of index numbers to its users are listed as under:

1. Index Numbers Help in Formulating Policies and Decision Making: Formulation of good policies for the future depends upon past trends. Behaviour of the index numbers is studied carefully before making any policies. Index numbers of the data relating to prices, production, profits, imports and exports, personnel and financial matters are indispensable for any organisation in efficient planning and formulation of executive decisions.

For example the cost of living index numbers help the employers in deciding about the increase in dearness allowance of their employees or adjusting their salaries and wages in accordance with changes in their cost of living.

2. Reveal Trends and Tendencies: The index numbers study the relative changes in the level of a phenomena. So, they would disclose the general trend for a variable or group of variables in time series data. For example by examining the index numbers of production (industrial and agricultural), volume of trade, imports and exports etc. for the last few years, we can draw useful conclusions about the trend of production and business activity.

3. Index Numbers Measure the Purchasing Power of Money: Index numbers are helpful in finding out the intrinsic value of money as contrasted with its nominal worth. The cost of living index numbers determine whether the real wages are rising or falling, money wages remaining unchanged.

For example, suppose that the cost of living index for any year, say, 2010 for a particular class of people with 2001 as base year is 200. If a person belonging to that class gets **Rs.** 300 in 2001, then in order to maintain the same standard of living as in 2001 (other factors remaining constant) his salary in 2010 should be $\frac{200}{100}$ x 300 = Rs.600.

4. Aid in Deflation: Index numbers are very useful for deflating (or adjusting) the original data. In obtaining real income from inflated income, real wages from nominal wages, and real sales from nominal sales and so on, the index numbers are immensely useful.

5. They are Economic Barometers. As described above various index numbers are computed for different purposes, say employment, trade, transport, agriculture, industry, etc., and these are of immense value in dealing with different economic issues. Like barometers which are used in Physics and Chemistry to measure atmospheric pressure, index numbers are rightly termed as 'economic barometers' or 'barometers of economic activity' which measure the pressure of economic and business behaviour.

In the words of G. Simpson and F. Kafka, "Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies."

6. Other Uses: Index numbers are thus, become indispensable tool in business planning and formulation of executive decisions of any concern and theses are today one of the most widely used statistical devices. Fairly good appraisal of trade and business activities of country can be done if careful study is carried out by constructing index. In brief the uses of index numbers are shown below:

- They measure the relative change
- They are of better comparison.
- They are wage adjuster.
- They compare standard of living.
- They are special type of averages.

Activity B:

1. "Index numbers are the devices for measuring differences in the magnitude of a group of related variables." Discuss this statement and point out the important uses of index numbers.

10.3 Types of Index Numbers

Index numbers may be divided into three categories:

(a) Price Index

(b) Quantity Index, and

(c) Value Index

(a) Price Index: It is the most commonly used index. It compares the prices of various commodities from one period to another. We may have steel price index, sugar price index, vegetables price index etc. For these purpose, a well known price index is called consumer price index (CPI), it is tabulated at regional or National level so as to establish the price levels of various consumer goods and services. This is an effective measure and useful index of cost of living. Index numbers are further divided into the following two categories:

(i) Wholesale Price Index Numbers: It reflect the changes in general price level of the country such as wholesale price index number prepared by government.

(ii) Retail Price Index Numbers: These index numbers show general changes in retail prices of various commodities such as consumption goods, stock and shares, bank deposits, consumer price index etc.

(b) Quantity Index: Quantity index numbers study the changes in the volume of goods produced, distributed or consumed, like the indices of agricultural production, industrial production, imports and exports, etc. These types of indices are useful for measuring the changes in level of physical output in an economy during some period compared with other period.

(c) Value Index: The last type of Index, the value index, measures changes in total monetary worth, during some period compared with other period. The value index numbers are intended to study the changes in the total value (quantity multiplied by price) of production.

10.4 Precautions in Construction of Index Numbers

Index numbers which are not properly compiled will, not only lead to wrong and fallacious conclusions but might also prove to be dangerous. So, the construction of index number requires a careful study of some aspects which often called 'precautions or problems'. The given below are some precautions:

1. Purpose of Index Numbers: It is essential to be clear about the purpose for which the index number is used. Every index number has its own particular uses. For example if it is used for measuring consumers' price, there is no need of including wholesale prices. Similarly, if it is employed for studying cost of living, there is no need of including the price of capital goods.

2. Selection of Base Period: Base period is the period against which comparisons are made. One has to be very careful in selecting a base period. If we select inappropriate base, then distortion can be large. Taking a year of large increase in prices due to abnormal causes may not be appropriate base. Also may be a large consumption of a commodity at the time of natural calamity may not reflect correct index level at the base year. Before selecting a base period the following points should be kept in mind:

(a) The base period should not get affected by extra-ordinary events like war, earthquakes, famines, booms, etc. it should be a normal one
(b) It should be relatively current i.e., it should not be too distant in the past because we are interested in the changes relating to the present period only.
(c) The base may be fixed, chain or average depending upon the purpose of constructing the index.

3. Selection of Commodities or Items: Only those commodities or items which are fairly represented and uniform quality should be selected for inclusion in the construction of index number. White selecting the sample the following points should be kept in mind:

- The commodities selected should be relevant to the purpose of the index.
- Select the adequate number of representative items from each group (Neither too small nor too large)
- Classify the whole relevant group of items or commodities into relatively homogeneous sub-groups.

4. Selection of Weights: 'Weights' imply the relative importance of the different variables. Due to

inappropriate importance given to various factors in calculation of index, the calculated value willbe found distorted and may not be the true representative of the decision variable. So proper weight should be assigned to different commodities with their relative importance in the group. Assigning weights can be done by (a) Implicit method; items are implicitly weighted, assumed to equal importance) and (b) Explicit method (weights are assigned either in quantity terms or in value terms). The choice of method of weighing depends on the purpose, scope, availability of data. Like in developing a composite index, such as the Consumer Price Index, we must consider changes in some variables to be more important than changes in others like, wheat should be given more importance compared to sugar.

5. Data Collection: The basic data used must be reliable, authentic and suitable for the purpose. Sufficient data is also required to arrive at a worthwhile information deduction. The source of data depends on their information requirements. In dealing with broad areas of national economy and the general level of business activity, publications such as the Federal Reserve Bulletin, Moody's, Monthly Labour Review, and the Consumer Price Index provide a wealth of data. Almost all government agencies distribute data about their activities, from which index numbers can be computed. Many financial newspapers and magazines provide information from which index numbers can be computed. When you read these sources, you will find that many of them use index numbers themselves.

6. Selection of Average: Averages play a vital role in arriving at a single index number summarizing a large volume of information. Arithmetic mean and geometric mean are used in it construction but theoretically, the geometric mean is preferred because it is less susceptible to variation; it gives equal weight to equal ratio of change.

7. Price Collection: After selecting the items, the next problem is to collect their prices. The price of a commodity varies from place to place and even from shop to shop in the same market. So, it is very difficult to consider and compile prices from every market, from every shop and for all periods. Therefore, we should select, a sample market, which are well known for trading in a particular commodity and also collect the data of price for that commodity from the agencies such as the Chambers of Commerce, News Correspondents etc. and further we should compare for validity and reliability.

8. Selection of Appropriate Formula: A large number of formulas have been devised for constructing the index numbers. A decision has, therefore, to be made as to which formula is the most suitable for the purpose. The choice of the formula depends upon the availability of the data regarding the prices and quantities of the selected commodities in the base and/or current year.

Activity C:

1. What is an Index number? Why index numbers are called "Economic barometers"?
2. Analyze the problems in the construction of index numbers and comment.

10.5 Notations and Methods of Constructing Index Numbers

Notations:

Base Year: The year selected for comparison i.e. the year with reference to which comparisons are made. It is denoted by '0'.

Current Year: The year for which comparisons are sought or required.

P_0 is price of a commodity in the base year,
P_1 is price of a commodity in the current year,
q_0 is quantity of a commodity in the base year,
q_1 is quantity of a commodity in the current year,
W is weight assigned to a commodity according to its relative importance in group,
P_{01} is price index number for the current year,

P_{10} is price index number for the base year,
Q_{01} is quantity index number for the current year, and
q_{10} is quantity index number for the base year

Methods of Constructing Index Numbers

(I) Price Index Numbers

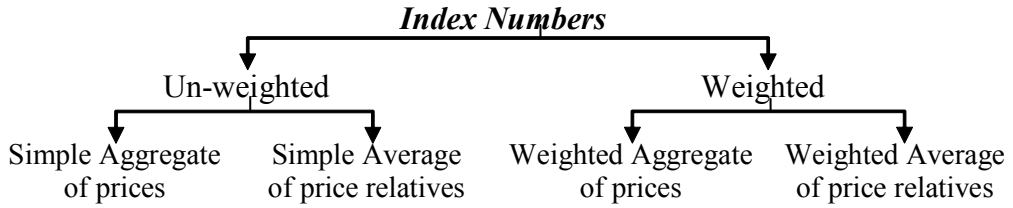
Methods of constructing index numbers can broadly be divided into two classes namely:

(A) Un-weighted Index Numbers, and
(B) Weighted Index Numbers.

In case of un-weighted indices, weights are not assigned, whereas in the weighted indices weights are assigned to the various items. Each of these types may be further classified under two heads:

(i) Aggregate of Prices Method, and
(ii) Average of Price Relatives Method.

The following chart illustrates the various methods:



(A) Un-weighted Index Numbers:

(i) Simple (Un-weighted) Aggregate Method: This is the simplest methods of constructing index numbers and consists in expressing the total price, i.e., aggregate of prices (of all the selected commodities) in the current year as a percentage of the aggregate of prices in the base year. Thus, the price index for the current year w.r.t. the base year is given by:

$$P_{10} = \frac{\Sigma p_1}{\Sigma p_0} x100$$

Where Σp_1 is the aggregate of prices in the current year and Σp_0 is the aggregate of prices in the base year.

This method suffers from a drawback that equal weight is given to all the items irrespective of their relative importance.

Illustration: 1

From the following data calculate Index Number by simple aggregate method.

Commodity	**:**	A	B	C	D
Price in 1990 (Rs.)	**:**	162	256	257	132
Price in 1991 (Rs.)	**:**	171	164	189	145

Solution:

Computation of Price Index Number

	Price (in Rupees)	
	1990 (P_0)	**1991(p_1)**
A	162	171
B	256	164
C	257	189
D	132	145
Total	**Σp_0=807**	**Σp_1=669**

The price index number using simple aggregate method is given by:

$$P_{10} = \frac{\Sigma p_1}{\Sigma p_0} X\ 100$$

$$= \frac{669}{807}\ X\ 100$$

$$= 82.90$$

Illustration: 2

Following prices are indicated for 2005 (base year) and for 2010 (the current year). Calculate the un-weighted aggregates price index for the data

Variables	Prices	
	2005	**2010**
Tomatoes (per kg.)	Rs. 15.00	Rs. 19.00
Egg (per dozen)	Rs. 20.00	Rs. 24.00
Petrol (Per Liter)	Rs. 22.50	Rs. 30.70
Juices (Per Liter)	Rs. 61.00	Rs. 69.00

Solution:

For computing price index (un-weighted) we have

Σp_1 = Rs. 19+24+30.70+69 = Rs. 142.70
Σp_0 = Rs. 15+20+22.50+61=Rs. 118.50

Un-weighted aggregates price index

$$= \frac{142.70}{118.50}\ \text{x}\ 100 = 120$$

(ii) Simple Average of Price Relatives: In order to calculate, we can compare the ratio of current prices to the base prices and then the index is calculated by multiplying the rate by 100. We, then, take the average of all the ratios summed up together. The general relationship now undergoes the change as follows:

Un-weighted average of relative price index

$$= \frac{\Sigma\ \frac{P_1}{P_0}}{n} x100$$

The simple average of price relatives method is superior to the simple aggregate of prices method in two respects:

(i) Since we are comparing price per liter with price per liter, and price per kilogram with price per kilogram the concealed weight due to use of different units is completely removed.
(ii) The index is not influenced by extreme items as, equal importance is given to all items.

The greatest drawback of unweighted indices is that equal importance or weight is given to all items included in the index number which is not proper. As such, unweighted indices are of little use in practice.

Illustration: 3

From the following data, construct index number by simple average of price relatives using arithmetic mean

Commodity	1998 Price (Rs.)	2001 (Price (Rs.)
Wheat	800/quintal	1000/quintal
Rice	15/Kg.	19/Kg.
Milk	12/Liter	15/Liter
Eggs	10/Dozen	12/Dozen
Sugar	14/Kg.	18/Kg.

Solution:

Commodity	P_0	P_1	P (Price relative)
Wheat	800	1000	1000 x 100/800 = 125
Rice	15	19	19 x 100/15 = 126.67
Milk	12	15	15 x 100/12 = 125
Eggs	10	12	12 x 100/10 = 120
Sugar	14	18	18 x 100/14 = 128.57
			625.24

Using simple Arithmetic Mean:

Average of relative price index

$$= \frac{\sum \frac{P_1}{P_0}}{n} x100$$

$$= \frac{625.24}{5} = 125.05$$

(B) Weighted Index Numbers:

The purpose of weighting is to make the index numbers more representative and to give more importance to them. Weighted index numbers are of two types. They are:

(i) Weighted Aggregate Index Numbers. According to this method, prices themselves are weighted by quantities; i.e., p x q. Thus physical quantities are used as weights. Here are various methods of assigning weights, and thus various formulas have been formed for the construction of index numbers. Some of the important formulae are given below:

1. Laspeyre's Method
2. Paasche's Method
3. Dorbish and Bowley's Method
4. Marshall-Edgeworth Method
5. Kelly's Method, and
6. Fisher's Ideal Method

1. Laspeyre's Method. In this method, base year quantities are taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} X\ 100$$

Where P_1 = Price in the current year

P_0 = Price in the base year

q_0 = Quantity in the base year

According to this method, the index number for each year is obtained in three steps:

(a) The price of each commodity in each year is multiplied by the base year quantity of that commodity. For the base year, each product is symbolised by P_0q_0, and for the current year by P_1q_0.

(b) The products for each year are totaled and $\sum P_1q_0$ and $\sum P_0q_0$ are obtained.

(c) $\sum P_1q_0$ is divided by $\sum P_0q_0$ and the quotient is multiplied by 100 to obtain the index.

Laspeyre's index is very widely used in practice. It tells us about the change in the aggregate value of base period list of goods when valued at given period price. However, this index has one drawback. It does not take into consideration the changes in the consumption pattern that take place with the passage of time.

2. **Paasche's Method**. In this method, the current year quantities are taken as weights: symbolically,

$$P_{01(Pa)} = \frac{\sum p_1q_1}{\sum p_0q_1} x100$$

According to this method, the index number for each year is obtained in following steps:

1. Multiply current year prices of various commodities with current year weights and obtain P_1q_1.

2. Multiply the base year prices of various commodities with the current year weights and obtain P_0q_1.

3. $\sum P_1q_1$ is divided by $\sum P_0q_1$ and the quotient is multiplied by 100 to obtain the index

Although this method takes into consideration the changes in the consumption pattern, the need for collecting data regarding quantities for each year or each period makes the method very expensive. Hence, where the number of commodities is large, Paasche's method is not used in practice.

3. Bowley Dorbish Method. This is an index number got by the arithmetic mean of Laspeyre's and Paasche's methods; symbolically (This method takes into account both the current and the base periods). Symbolically

$$P_{01(B)} = \frac{\frac{\sum p_1q_0}{\sum p_0q_0} + \frac{\sum p_1q_1}{\sum p_0q_1}}{2} x100 = \frac{L+P}{2}$$

L = Laspeyre's method
P = Paasche's method

4. Marshall-Edgeworth Method. In this method, the totals of base year and current year quantities are taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\sum P_1(q_0+q_1)}{\sum P_0(q_0+q_1)} x100$$

$$or\ P_{01} = \frac{\sum p_1q_0}{\sum P_0q_0} + \frac{\sum P_1q_1}{\sum p_0q_1} x100$$

5. Kelly's Price Index or Fixed Weights Index. This formula, named after Truman L. Kelly, requires the weights to be fixed for all periods and is also sometimes known as *aggregative index with fixed weights* and is given by the formula:

$$P_{01} = \frac{\sum p_1q}{\sum P_0q} X\ 100$$

Where the weights are the quantities (q) which may refer to some period (not necessarily the base year or

the current year) and are kept constant for all periods. The average (A.M. or G.M.) of the quantities consumed of two, three or more years may be used as weights.

Kelly's fixed base index has a distinct advantage over Laspeyre's index because unlike Laspeyre's index the change in the base year does not necessitate a corresponding change in the weights which can be kept constant until new data become available to revise the index. As such, currently this index is finding great favour and becoming quite popular.

6. Fisher's Ideal Index. This method is the geometric mean of Laspeyre's and Paasche's indices. The formula for constructing the index is :

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} x \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$$

Because of the following reasons, Fisher's formula is known as 'ideal':

• It takes into account prices and quantities of both current year as well as base year.
• It uses geometric mean which, theoretically, is the best average for constructing index numbers.
• It satisfies both time reversal test and the factor reversal test.
• It is free from bias. The weight biases embodied in Laspeyre's and Paasche's methods are crossed geometrically and thus eliminated completely.

Illustration: 4

By using Laspeyre's method, calculate the weighted price index for the year 2010 when the given data indicates the prices and consumption levels of various commodities.

Commodities	*Base Price (2007)*	*Current Price (2010)*	*Average quantity consumed (2007)*
Potatos (per kg)	Rs.5.10	Rs.4.50	4000 Kgs
Milk (per litre)	Rs.14.00	Rs.17.00	800 litres
Eggs (per doz)	Rs.21.00	Rs.24.00	2000 dozens
Bread (per loaf)	Rs.17.50	Rs.19.00	350 loaves

Solution:

For working out Laspeyre's Price Index, we prepare the table as follows:

Commodities	*Price in 2007*	*Price in 2010*	*Quantity in 2007*		
	(P_0)	*(P_1)*	*(Q_0)*	*P_0Q_0*	*P_1Q_0*
Potatos (per kg)	Rs.5.10	Rs.4.50	4000 Kgs	20,400	18,000
Milk (per litre)	Rs.14.00	Rs.17.00	800 litres	11,200	13,600
Eggs (per doz)	Rs.21.00	Rs.24.00	2000 dozens	42,000	48,000
Bread (per loaf)	Rs.17.50	Rs.19.00	350 loaves	6,125	6,650

From the above calculations

ΣP_0Q_0 = Rs. 79,725

ΣP_1Q_0 = Rs. 86,250

Laspeyre's Price Index = = $\frac{\sum P_1 Q_0}{\sum P_0 Q_0}$ x 100

$$= \frac{86,250}{79,725} x100$$

$$= 108$$

Illustration: 5

From the following data, construct the Laspeyre's, Paasche's and Fisher's indices of prices:

Item	*Base Year*		*Current Year*	
	P_0	q_0	P_1	q_1
A	4	20	10	15
B	8	4	16	5
C	2	10	4	12
D	10	5	20	6

Solution:

Calculation of Price Index Numbers

Item	*Base Year*		*Current Year*					
	P_0	q_0	P_1	q_1	P_0q_0	P_1q_0	P_0q_1	P_1q_1
A	4	20	10	15	80	200	60	150
B	8	4	16	5	32	64	40	80
C	2	10	4	12	20	40	24	48
D	10	5	20	6	50	100	60	120
Total					**ΣP_0q_0=182**	**ΣP_1q_0=404**	**ΣP_0q_1=184**	**ΣP_1q_1=398**

Laspeyre's Price Index (L) $= \frac{\sum P_1q_0}{\sum P_0q_0} \times 100 = \frac{404}{182} x100 = 221.98$

Paasche's Method (P) $= \frac{\sum P_1q_1}{\sum P_0q_1} \times 100 = \frac{398}{184} x100 = 216.30$

Fisher's Ideal Index $= \sqrt{L\ x\ P}$

$= \sqrt{221.98\ x\ 216.3} = 219.12$

Illustration: 6

From the following data, calculate the price index numbers for 2008 with 2000 as base by:

(a) Laspeyre's method

(b) Paasche's method

(c) Bowley method

(d) Marshall-Edgeworth method

(e) Fisher's Ideal method

Item	*2000*		*2008*	
	Price (Rs.)	*Quantity (unit)*	*Price (Rs.)*	*Quantity (unit)*
Maize	70	28	140	21.0
Millet	175	35	210	17.5
Sugar	140	52.5	175	52.5
Coconut	70	70.0	70	87.5

Solution:

Calculation of Price Index Numbers

Item	P_0	q_0	P_1	q_1	P_1q_0	P_0q_0	P_1q_1	P_0q_1
Maize	70	28	140	21.0	3920	1960	2940	1470
Millet	175	35	210	17.5	7350	6125	3675	3062.5
Sugar	140	52.5	175	52.5	9187.5	7350	9187.5	7350
Coconut	70	70	70	87.5	4900	4900	6125	6125
					25357.5	**20335**	**21927.5**	**18007.5**

(a) Laspeyre's method $= P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} x100 = \frac{25357.5}{20335} x100 = 124.69$

(b) Paasche's method $= P_{01} = \frac{\sum p_1q_1}{\sum p_0q_1} x100 = \frac{21927.5}{18007.5} x100 = 121.76$

(c) Bowley method $= P_{01} = \frac{L+P}{2} = \frac{124.69 + 121.76}{2} = 123.22$

(d) Marshall-Edgeworth method $= P_{01} = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} x100$

$= \frac{25357.5 + 21927.5}{20335 + 18007.5} x100 = 123.32$

(e) Fisher's Ideal method $= P_{01} = \sqrt{L\ x\ P}\ x100$

$= \sqrt{124.70\ x\ 121.77}\ x100 = 123.23$

Activity D:

From the data given below, construct index number of prices for 2006 with 2000 as base, using (i) Laspeyre's method, (ii) Paasche's method, (iii) Bowley-Drobisch method, (iv) Marshall - Edgeworth method, and (v) Fisher's ideal formula.

Commodity	2000		2006	
	Price per unit	Expenditure in rupees	Price per unit	Expenditure in rupees
A	2	10	4	16
B	3	12	6	18
C	1	8	2	14
D	4	20	8	32

(ii) Weighted Average of Price Relatives: This method is similar to the simple average of price relatives method with the fundamental difference that explicit weights are assigned to each commodity included in the index. Since price relatives are in percentages, the weights used are value weights. The following steps are taken in the construction of weighted average of price relatives index:

(i) Calculate the price relatives, $\left[\frac{P_1}{P_0} x100\right]$, for each commodity

(ii) Determine the value weight of each commodity in the group by multiplying its price in base year by its quantity in the base year, i.e., calculate P_0q_0 for each commodity. If, however, current year quantities are given, then the weights shall be represented by P_1q_1.

(iii) Multiply the price relative of each commodity by its value weight as calculated in above (ii).

(iv) Total the products obtained under (iii) above.

(v) Divide the total (iv) above by the total of the value weights. Symbolically index number obtained by the method of weighted average of price relatives is:

$$P_{01} = \frac{\sum\left[\left(\frac{P_1}{P_0}x100\right)P_0q_0\right]}{\sum P_0q_0} \; or \; \frac{\sum PV}{\sum V}$$

Illustration: 7

Based on the data given in illustration 4 calculate the weighted average of relatives index.

Solution:

Calculation of Weighted Average of Relatives Index

Commodities	(P_0)	(P_1)	(Q_0)	$\left[\frac{P_1}{P_0}\right]x100$	P_0Q_0	$\left[\frac{P_1}{P_0}\right]x100\ (P_0Q_0)$
Potatos	Rs.5.10	Rs.4.50	4000	88.23	20,400	17,99,892
Milk	Rs.14.00	Rs.19.00	800	135.71	11,200	15,19,952
Eggs	Rs.21.00	Rs.24.00	2000	114.28	42,000	47,99,760
Bread	Rs.17.50	Rs.19.00	350	108.57	6,125	6,64,991

From the above calculations, we obtain the values

ΣP_0Q_0 = Rs.79,725

and $\sum\left[\left(\frac{P_1}{P_0}\right)x100(P_0Q_0)\right] = 87,84,595$

$\therefore$ Weighted Average of the relatives index

$$= \frac{\sum\left[\left(\frac{P_1}{P_0}\right)x100(P_0Q_0)\right]}{\sum P_0Q_0}$$

$$= \frac{87,87,595}{79,725} = 110$$

(II) Quantity Index Numbers:

A quantity index number is a statistical device which measures changes in quantities in current year as compared to base year. Quantity index numbers reflect the relative changes in the quantity or volume of goods produced, consumed, marketed or distributed in any given year *w.r.t.* to some base year.

The formulae for calculating the quantity index numbers can be directly written from price index numbers simply by interchanging the role of price and quantity.

Thus quantity index by different methods is:

(a) Laspeyre's method = $Q_{01} = \frac{\sum q_1p_0}{\sum q_0p_0}x100$

(b) Paasche's method = $Q_{01} = \frac{\sum q_1p_1}{\sum q_0p_1}x100$

(c) Fisher's Ideal method = $Q_{01} = \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} x \frac{\sum q_1 P_1}{\sum q_0 P_1}}\ x100$

Illustration: 8. Compute quantity index for the year 1992 with base 1990=100, for the following data, using (i) Laspeyre's method (ii) Paasche's method. (iii) Fisher's ideal formula.

Item	*Price*		*Quantities*	
	1990	*1992*	*1990*	*1992*
A	5.00	6.50	5	7
B	7.75	8.80	6	10
C	9.63	7.75	4	6
D	12.50	12.75	9	9

Solution:

Computation of Quantity Index

Commodity	P_0	q_0	P_1	q_1	q_0P_0	q_0P_1	q_1P_0	q_1P_1
A	5.00	5	6.50	7	25.00	32.50	35.00	45.50
B	7.75	6	8.80	10	46.50	52.80	77.50	88.00
C	9.63	4	7.75	6	38.52	31.00	57.78	46.50
D	12.50	9	12.75	9	112.50	114.75	112.50	114.75
					Σq₀P₀ = 222.52	***Σq₀P₁ = 231.05***	***Σq₁P₀ = 282.78***	***Σq₁P₁ = 294.75***

(i) Laspeyre's quantity index or Q_{01} $= \frac{\sum q_1 P_0}{\sum q_0 P_0} x100$

$$= \frac{282.78}{222.52} x100 = 127.08$$

(ii) Paasche's quantity index or Q_{01} $= \frac{\sum q_1 P_1}{\sum q_0 P_1} x100$

$$= \frac{294.75}{231.05} x100 = 127.57$$

(iii) Fisher's quantity index or Q_{01} $= \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} x \frac{\sum q_1 P_1}{\sum q_0 P_1}}\ x100$

$$= \sqrt{\frac{282.78}{222.52} x \frac{294.75}{231.05}}\ x100$$

= 1.273 x 100
= 127.3

(III) Value Index Numbers:

These index measures the changes in the total value of the variable. Since value is a combination of price and quantity, it can be called a composite index. The only negative issue is that composite value index does not distinguish the variations in individual values of price or quantity separately. Value index numbers are obtained on expressing the total value (or expenditure) in any given year as a percentage of the same in the base year. Symbolically, we write

$$V_{01} = \frac{Total\ value\ in\ current\ year}{Total\ value\ in\ base\ year} x100 \quad \Rightarrow \quad V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} x100$$

The value index number based on the information given in illustration 8 can be calculated as under:

Value Index number $= V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} x100 = \frac{294.75}{222.52} x100 = 132.46$

10.6 Test of Consistency

As there are several formulae for constructing index numbers the problem is to select the most appropriate formula in a given situation. Prof. Irving Fisher has suggested two tests for selecting an appropriate formula. These are:

(A) Time Reversal Test, and

(B) Factor Reversal Test

(A) Time Reversal Test:

This test requires that the formulae for calculating an index number should give consistent results in both the directions, i.e. forward and backward. Or in other words, the index of period 1 with period 0 base should be reciprocal of the index of period 0 with period 1 as base i.e. $P_{01} = 1/P_{10}$ or $p_{01} \times P_{10} = 1$.

This test is satisfied by the Fisher's Ideal Index.

We can write

$$P_{01}^{F1} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} x \frac{\sum p_1 q_1}{\sum p_0 q_1}} \ (dropping\ 100)$$

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} x \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$\therefore \quad P_{01}^{F} x P_{10}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} x \frac{\sum p_1 q_1}{\sum p_0 q_1} x \frac{\sum p_0 q_1}{\sum p_1 q_1} x \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{1} = 1$$

(B) Factor Reversal Test

This test requires that the product of price index and the corresponding quantity index numbers should be equal to the value index number i.e. $P_{01} x Q_{01} = V_{01}$.

This test is also satisfied by the Fisher's ideal index.

We can write

$$P_{01}^{F1} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} x \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

$$\text{and } Q_{01}^{F1} = \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} x \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$\therefore \quad P_{01}^{F} x q_{01}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} x \frac{\sum p_1 q_1}{\sum p_0 q_1} x \frac{\sum q_1 p_0}{\sum q_0 p_0} x \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \sqrt{\left(\frac{\sum p_1 q_1}{\sum p_0 q_0}\right)^2} = \frac{\sum p_1 q_1}{\sum p_0 q_1} = V_{01}$$

Illustration: 9

Compute Index Number, using Fishers Ideal formula and show that it satisfies time-reversal test and factor-reversal test.

	Quantity	*Base Year Price*	*Quantity*	*Current year Price*
A	12	10	15	12
B	15	7	20	5
C	24	5	20	9
D	5	16	5	14

Solution:

Computation of Index Number

Commodity	q_0	p_0	q_1	p_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	12	10	15	12	144	120	180	150
B	15	7	20	5	75	105	100	140
C	24	5	20	9	216	120	180	100
D	5	16	5	14	70	80	70	80
					Σp₁q₀ = 505	***Σp₀q₀ = 425***	***Σp₁q₁ = 530***	***Σp₀q₁ = 470***

$$P_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} x \frac{\sum p_1q_1}{\sum p_0q_1}} \text{ x } 100$$

$$= \sqrt{\frac{505}{425} x \frac{530}{470}} \text{ x } 100$$

$$= \sqrt{1.188 \; x \; 1.128} \; x \; 100$$

$$= \sqrt{1.340} \; x \; 100 = 1.158 \; x \; 100$$

$$= 115.8$$

(a) Time-Reversal Test

Time Reversal Test is satisfied when $P_{01} \times P_{10} = 1$

$$P_{10} = \sqrt{\frac{\sum p_0q_1}{\sum p_1q_1} x \frac{\sum p_0q_0}{\sum p_1q_0}}$$

$$= \sqrt{\frac{470}{530} x \frac{425}{505}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{505}{425} \; x \; \frac{530}{470} \; x \; \frac{470}{530} x \frac{425}{505}} = \sqrt{1}$$

$= 1$

(b) Factor-Reversal Test

Factor-Reversal test is satisfied when

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} x \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} x \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$P_{01} x Q_{01} = \sqrt{\frac{505}{425} x \frac{530}{470} x \frac{470}{425} x \frac{530}{505}}$$

$$= \sqrt{\frac{530}{425}} \; i.e., \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

$$P_{01} x Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Hence the given data satisfies the time-reversal test and factor-reversal test.

Activity E:

1. What are the tests of a good index number? Define Fisher's Ideal index number and show that it satisfies all these tests.
2. State whether the following statements are 'True' or 'Untrue':
(i) Arithmetic mean is the most appropriate average for constructing the index numbers.
(ii) Paasche's Index number is based on base year quantity.
(iii) Fisher's Index Number is an Ideal Index Number.
(iv) Fisher's Index Numbers is the arithmetic average of Laspeyre's and Paasche's Index Numbers.
(v) Time reversal test is satisfied by both formulas Fisher and Kelly's.

10.7 Fixed and Chain Base Index Numbers

Fixed Base Index Numbers:

When the comparison of (prices or quantities etc.) various periods are done with reference to a particular or fixed period, we get an index number series with fixed base.

Chain Base Index Numbers:

The main problem with a fixed base series arises when the current year becomes too far off from the base year. In such a situation, it may happen that the commodities which used to be very important in the base year are no longer so in current year. Furthermore, certain new commodities might be in use while some old commodities are dropped in current year. This problem is often solved by constructing Chain Base Index Numbers. A chain base index number is an index number with previous year as base.

Differences between Chain Base Method and Fixed Base Method

S. No.	*Chain Base*	*Fixed Base*
1.	The base year changes.	The base year does not change.
2.	Here the link relative method is used.	No such link relative method is used.
3.	Introduction and deletion of items are easy to calculate, without recalculation of the entire series.	Any change in the commodities, will involve the entire index number to be recast.
4.	The calculations are tedious	The calculations are simple.
5.	It is difficult to understand.	It is simple to understand.

6.	It cannot be computed if data for any one year are missing.	There is no such problem.
7.	It is suitable for short period only.	It is suitable for long periods only.
8.	Weights can be adjusted as frequently as possible.	Weights cannot be adjusted so frequently.
9.	Index number is wrong if an error is committed in the calculation of any link index number.	This is not so, the error is confined to the index of that year only.

Illustration: 10. From the following data relating to the wholesale prices of wheat for six years, construct index numbers using (a) 1990 as base, and (b) by chain base method.

Year	*Price (per quintal) Rs.*	*Year*	*Price (per quintal) Rs.*
1990	100	**1993**	130
1991	120	**1994**	140
1992	125	**1995**	150

Solution:

(a) Computation of Index numbers with 1980 as base:

Year	*Price of wheat*	*Index Number 1980 = 100*	*Year*	*Price of wheat*	*Index Number 1980 = 100*
1990	100	100	**1993**	130	$\frac{130}{100}x100 = 130$
1991	120	$\frac{120}{100}x100 = 120$	**1994**	140	$\frac{140}{100}x100 = 140$
1992	125	$\frac{125}{100}x100 = 125$	**1955**	150	$\frac{150}{100}x100 = 150$

(b) Construction of Link Relative Indices

Year	*Price of wheat*	*Link Relative Index*	*Year*	*Price of wheat*	*Link Relative Index*
1990	100	100	**1993**	130	$\frac{130}{125}x100 = 104$
1991	120	$\frac{120}{100}x100 = 120$	**1994**	140	$\frac{140}{130}x100 = 107.692$
1992	125	$\frac{125}{120}x100 = 104.167$	**1995**	150	$\frac{150}{140}x100 = 107.14$

Conversion of Link Relatives into Chain Relatives:

Chain relatives or chain indices can be obtained either directly or by converting link relatives into chain relatives with the help of the following formula:

$$\text{Chain relative for current year} = \frac{\text{Link Relative for the current year} \times \text{Chain relative for the previous year}}{100}$$

Taking the data from illustration 10, we can show the method of conversion as follows:

Year	*Price of wheat*	*Link relative*	*Chain relative*
1990	100	100.00	100
1991	120	120.00	$\frac{120\ x\ 100}{100} = 120$
1992	125	104.167	$\frac{104.167\ x\ 120}{100} = 125$
1993	130	104.00	$\frac{104\ x\ 125}{100} = 130$
1944	140	107.692	$\frac{107.692\ x\ 130}{100} = 140$
1955	150	107.14	$\frac{107.14\ x\ 140}{100} = 150$

10.8 Base Shifting, Spicing and Deflating the Index Numbers

Base Shifting:

Sometimes it becomes necessary to shift the base from one period to another. This becomes necessary either because the previous base has become too old and has become useless for comparison purposes or because comparison has to be made with another series of index numbers having different base period.

The following formula must be used in this method of base shifting:

Index Number (based on New Base Year)

$$= \frac{Current\ year's\ old\ index\ number}{New\ base\ year's\ old\ index\ number} x\ 100$$

Illustration: 11

Shift the base of the following series to 1977.

Year	1995	1996	1997	1998	1999	2000
Index No.	125	155	185	220	265	320

Solution:

To shift the base at 1987, we multiply every index number by 100/185.

Year	1995	1996	1997	1998	1999	2000
Index No.	67.6	83.8	100	118.9	143.2	173.0

Splicing Two Index Number Series:

The statistical method connects an old index number series with a revised series in order to make the series continuous is called splicing. The articles which are included in an index number may become out of fashion or go out of the market. New articles come into the market, for which relative importance may also change. So it is necessary to include the articles in the index number. The old series of index number is discontinued and we must construct a new series and must take the year of discontinuation as the first base.

Thus we connect the new set of index with the old discontinued one. The formula is:

$$= \frac{Index\ no.of\ current\ year\ x\ old\ Index\ No.of\ New\ base\ year}{100}$$

Illustration: 12

Two sets of Indices, one with 1986 as base and the other with 1994 as base are given below :

(a) Year	*Index Numbers*	*(b) Year*	*Index Number*
1986	100	**1994**	100
1987	110	**1995**	105
1988	120	**1996**	90
1989	190	**1997**	95
1990	300	**1998**	102
1991	330	**1999**	110
1992	360	**2000**	96
1993	390		
1994	400		

The Index (a) with 1986 base was discontinued in 1994. You are required to splice the second index number (b) with 1994 base to the first index number.

Solution:

Splicing of Index Numbers

Year	*Index Number (a) with 1976 as base*	*Index Number (b) with 1984 as base*	*Index Number (b) spliced to (a) with 1976 as base*
1986	100		
1987	110		
1988	120		
1989	190		
1990	300		
1991	330		
1992	360		
1993	390		
1994	400	100	$100\ x\frac{400}{100} = 400$
1995		105	$105\ x\frac{400}{100} = 420$
1996		90	$90\ x\frac{400}{100} = 360$
1997		95	$95\ x\frac{400}{100} = 380$
1998		102	$102\ x\frac{400}{100} = 408$
1999		110	$110\ x\frac{400}{100} = 440$
2000		96	$96\ x\frac{400}{100} = 384$

Deflating:

Deflating is the process of making allowances for the effect of changing price levels. With increasing price levels, the purchasing power of money is reduced. As a result, the real wage figures are reduced and the real wages become less than the money wages. To get the real wage figure, the money wage figure may be reduced to the extent the price level has raised. The process of calculating the real wages by applying index numbers to the money wages so as to allow for the change in the price level is called deflating. Thus, deflating is the process by which a series of money wages or incomes can be corrected for price changes to find out the level of real wages or incomes. This is done with the help of the following formula:

Real Wage $= \frac{Money\ Wage}{Price\ Index}\ x\ 100, and$

Real Wage Index $= \frac{Real\ wages\ for\ the\ year}{Real\ wages\ for\ the\ Base\ year} x\ 100$

Illustration: 13

Given the following data:

Year	*Weekly take-home pay (wages)*	*Consumer Price Index*
1991	109.5	112.8
1992	112.2	118.2
1993	116.4	127.4
1994	125.08	138.2
1995	135.4	143.5
1996	138.1	149.8

(1) What was the real average weekly wage for each year?

(2) In which year did the employee have the greatest buying power?

(3) What percentage increase in the weekly wages for the year 1996 is required, if any, to provide the same buying power that the employees enjoyed in the year in which they had the highest real wages?

Solution:

Calculation of Real Wages

Year	*Weekly take-home pay (Rs.)*	*Consumer price Index*	*Real Wages*
1991	109.5	112.8	$\frac{109.5}{112.8} x\ 100 = 97.07$
1992	112.2	118.2	$\frac{112.2}{118.2} x\ 100 = 94.92$
1993	116.4	127.4	$\frac{116.4}{127.4} x\ 100 = 91.37$
1994	125.08	138.2	$\frac{125.08}{138.2} x\ 100 = 90.51$
1995	135.4	143.5	$\frac{135.4}{143.5} x\ 100 = 94.36$
1996	138.1	149.8	$\frac{138.1}{149.8} x\ 100 = 92.19$

(1) Real average weekly wage can be obtained by the formula:

Real Wage $= \frac{Money\ Wage}{Price\ Index} x\ 100$

(2) The employee had the greatest buying power in 1991 as the real wage was maximum in 1991.

(3) Absolute difference = 97.07 - 92.19 = + 4.88

10.9 Limitations of Index Numbers

Even though index numbers are very important in business and economic activities, they have their own limitations; they are:

1. The index numbers are only approximate indicators because their construction based on the sample data, and may not exactly represent the true changes in relative level of a phenomenon.

2. An index number does not take into account the quality of items.

3. Likelihood of error is possible at each stage of construction of index numbers, viz., (i) selection of commodities, (ii) selection of the base period, (iii) collection of data - prices and quantities of commodities, (iv) choice of formula - the procedure of weight age to be given.

4. Index number is an average and as such it suffers from all the limitations of an average.

5. Consumption is the result of taste, custom, attitude, etc., which are dynamic.

6. There is no unique index number that is acceptable to all.

7. There may be possibility of manipulation of the base year, price, commodities and quantity quotations in order to get the required results by the selfish persons.

10.10 Summary

Index numbers help to get an idea of the present day situation with regard to changes in production, consumption, exports and imports, national income, business level, cost of living, the price of a particular commodity or a group of commodities, Industrial or agricultural production, stocks and shares, sales and profits of a business house, volume of trade, factory production, wage structure of workers in various sectors, bank deposits, foreign exchange reserves, and so on.

Index numbers help in formulating policies and decision making, reveal trends and tendencies, measure the purchasing power of money, aid in deflation, and work like economic barometers. Index numbers may be divided into three categories price index, quantity index, and value index.

Construction of index numbers requires a careful study of some aspects like purpose of index numbers, selection of base period, selection of commodities or items, selection of weights, collection of data, selection of average, price collection, and selection of appropriate formula to construct the index numbers. Methods of constructing index numbers can broadly be divided into two classes namely: un-weighted index numbers, and weighted index numbers. In case of un-weighted indices, weights are not assigned, whereas in the weighted indices weights are assigned to the various items. Each of these types may be further classified under two heads: aggregate of prices method, and average of price relatives method.

Even though index numbers are very important in business and economic activities, they have their own limitations; like index numbers are based on the sample data, and may not exactly represent the true changes in relative level of a phenomena, error is possible at each stage of construction of index numbers, no unique index number is acceptable to all, and there is a possibility of manipulation of the base year, price, commodities and quantity quotations.

To conclude we can say that Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies

10.11 Key Words

Index Number: A ratio of the variable value at the current level to the base level, i.e. the ratio of variable change over a period.

Un-weighted Aggregates Index: Using all the values or data collected for study and allocating same importance to all the values.

Un-weighted Average of Relatives Method: Working out the index number by dividing the present or current level of the variable to its base value, multiplied by 100 and then dividing the summation of percentage values by the number of products to result in average value.

Weighted Aggregates Index: Using all the values considered, but assigning importance or weights to individual ratios.

Weighted Average of Relatives Method: Constructing the index number by allotting weight ages to values of each element in the composite.

Consumer Price Index: Indicates the variations in the prices of a given set of consumer items prepared either at regional level or at National level.

Fixed Base Method: When weight or importance to the variable is allotted based on a given specified fixed period.

Price index: Compares levels of prices from one period (the current) to another (the base) period.

Quantity index: An index comparing the quantity of the variable during a given period by time.

10.12 Self Assessment Questions

1. "Index numbers are economic barometers". Explain this statement and mention what precautions should be taken in making use of any published index numbers.
2. Define Index Number and mention its uses.
3. Define index numbers. Distinguish between fixed base and chain base method of constructing index numbers.
4. What are index numbers? What purpose do they serve? Discuss the various problems faced in the construction of index numbers.
5. State and explain Fisher's Ideal Formula for Price Index Number and why is it called Ideal?
6. Write explanatory notes on the following:

   (a) Deflating;
   (b) Splicing
   (c) Base Shifting;
7. Discuss briefly the uses and limitations of index numbers of prices.
8. Compute price index for the following data by (i) simple aggregative method, and (ii) average of price relative method by using arithmetic mean.

Commodities	**A**	**B**	**C**	**D**	**E**	**F**
Price 2003 (Rs.)	20	30	10	25	40	50
Prices 2004 (Rs.)	25	30	15	35	45	55

Ans. (i) 117.14, (ii) 122.9 (by A. M.)

9. Calculate weighted aggregative price index number taking 2001 as base, from the following data:

(Commodity)	**(Quantity consumed)**	**Price per Unit**	
		Base year 2001	**Current year 2004**
Wheat	4 Qtl.	80	200
Rice	1 Qtl.	120	250
Gram	1 Qtl.	100	150
Pulses	2 Qtl.	200	300

Ans. Weighted Index Number = 191.49

10. From the data given in the following table, calculate consumer's price index numbers for the year 2004 taking 2003 as base using (i) simple average, and (ii) weighted average of price relatives:

Items	Unit	Price in Rs.		Weight
		2003	2004	
Wheat	Kg.	0.50	0.75	2
Milk	Liter	0.60	0.75	5
Egg	Dozen	2.00	2.40	4
Sugar	Kg.	1.80	2.10	8
Shoes	Pair	8.00	10.00	1

Ans. (i) 127.34, (ii) 123.3

11. Find out the index number for the year 2004 from the following data using the weighted average of price relatives method:

Commodity	Weight	Price	
		2000	**2004**
Wheat	4	50	100
Milk	3	30	90
Egg	5	20	10
Sugar	3	60	90
Shoes	5	20	120

Ans.: 270

12. From the following data, calculate Fisher's Ideal Index:

Items	Price per unit (Rs.)		Quantity used	
	2003	**2004**	**2003**	**2004**
A	9.25	15.00	5	5
B	8.00	12.00	10	11
C	4.00	5.00	6	6
D	1.00	1.25	4	8

Ans.: 148.78

13. Given below are two series of index numbers, one based on 1997 and the other on 2000. Splice the new series on 1997 base:

Year	1997	1998	1999	2000	2001	2002	2003	2004
Old Series (A)	100	110	125	150	-	-	-	-
New Series (B)	-	-	-	100	105	120	130	150

Ans.: 100, 110, 125, 150, 157.5, 180, 195, 225

14. From the data given below construct index number of quantities, and of prices for 1970 with 1966 as base using (i) Laspeyre's formula, (ii) Paasche's formula, and (iii) Fisher's Ideal formula.

Commodity	1966		1970	
	Price (Rs.)	Quantity (Units)	Price (Rs.)	Quantity (Units)
A	5.20	100	6	150
B	4.00	80	5	100
C	2.50	60	5	72
D	12.00	30	9	33

Ans. (130.07, 131.02, 130.54) and for price (116.29, 117.14, 116.7)

15. From the following data calculate index numbers of real wages with 1999 as the base:

Year	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**
Average Wages (Rs.)	2400	2640	2860	3000	3420	4000	4200
Consumer Price Index	100	120	130	150	190	200	210

Ans. 120, 110, 110, 100, 90, 100, 100

16. From the data given below, calculate the price index by Fisher's ideal formula and then verify that Fisher's ideal formula satisfies both time reversal test and factor reversal test.

Commodity	*Base year*		*Current year*	
	Price (Rs.)	**Quantity ('000 tonnes)**	**Price (Rs.)**	**Quantity ('000 tonnes)**
A	56	71	50	26
B	32	107	30	83
C	41	62	28	48

Ans. (84.92)

10.13 Reference Books

1. Richard I. Levin and David S. Rubin, Statistics for Management
2. Gupta, S. P., Statistical Methods
3. Yadav, Jain, Mittal, Statistical Methods.
4. Nagar, K. N., Statistical Methods.
5. Gupta, C.B. and Gupta, Vijay, An Introduction to Statistical Methods.

Unit - 11 Correlation Analysis

Structure of Unit:

11.0 Objectives

After completing this unit, you will be able to : -

- Understand the meaning of Correlation.
- Explain various types of Correlation.
- Calculate correlation with the help of many graphic and mathematical methods.
- Interpret the significance of correlation with the help of probable error.
- Understand the utility of cause and effect relationship in correlation study.
- Evaluate the difference and importance of different methods.

11.1 Introduction and Definition of Correlation

The statistical methods so far discussed and analysed relate to one variable and throw light on the construction and shape of the univariate distribution. But sometimes, two or more such series are required to be analysed which changes simultaneously. In these bivariate distributions, we have to face some new questions, like does there exist association between the two variables ? If the value of one variable increases (or decreases) does it affect the other ? If yes, to what extent and in which direction? For example, we can put together the data of income and expenditure and analyse them, we may find that the expenditure increases with the increase of income and the expenditure decreases with the decrease of income. Generally upto a certain age, increase in age is associated with the increase in height of a child. Similarly, if we analyse the data of price and demand of a particular commodity, we may find that rise in the price of a commodity reduces its demand and vise-versa. Thus, a number of examples come across in our life in which two variables are inter dependent. In this unit the methods which are used in studying the relationship between two variables are being discussed.

Definition of Correlation

Correlation means a relation between two series or groups of data. If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in the others, then they are said to be correlated.

The statistical tool with the help of which the relationship between two or more than two variables is discovered and measured is called correlation.

According to Prof W. I. King, "If it is proved that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions we consider that the fact is established and that a relationship exists. This relationship is called correlation."

According to Prof. A. L. Boddington - "Whenever some definite correlation exists between two or more groups, classes or series of data, that is said to be a correlation.

E. Davenport has expressed his views as - "The whole subject of correlation refers to the inter-relation between separate characters by which they tend, in some degree, at least to move together."

According to A. M. Tuttle - "Analysis of co-variation of two or more variables is usually called correlation.

Thus two or more variables are said to be correlated if the change in one variable results or associated with a corresponding change in other variable.

Correlation and Causation

Correlation analysis helps us in determining the degree of relationship between two or more variables. But high degree of correlation doesn't always means of existence of cause and effect relationship.

When the changes in one series are the causes and the changes in the other series are the effects of the causes of the changes in first series, such a correlation will be significant.

There is a great difference between Association and Causation. When two variables are associated, it does not mean that they must have cause and effect relationship. Though cause and effect relationships always denotes correlation. On the basis of association, the calculation of correlation leads to illusory conclusions. Such relation is called 'Spurious correlation' or 'Nonsense Correlation'.

Following are the reasons of such correlation : -

(i) When high level correlation exists between two variables but nothing such exists in practice, such a correlation is called spurious, for example, despite high degree relation between increase in income and increase in blood pressure, it is spurious correlation because there is lack of any casual relation between the two variables.

(ii) When correlation in two or more variables definitely exists but it is difficult to decide which variable is 'cause' and which is 'effect' e.g., increase in sale of hot drinks (like tea, coffee) may correspond to increase in sale of woolens. Similarly despite correlation between bank deposits and loans it is difficult to know whether increase in deposits leads to increase in loans or vice-versa.

(iii) When both variables are independent but are affected equally due to some other cause, the correlation between them is nonsense, e.g., increase in the number of two wheelers leading to increase in new telephone connections is not directly connected but increase in income of people leads to increase in both variables. Similarly, increase or decrease in the yield of wheat and in the yield of rice. Both may be affected by rainfall. So correlation between the yields of both the crops is not because of causation.

Thus it is clear that causal relation is not essential to find correlation between two variables. On the other hand, it is essential to know cause and effect relation to avoid non-sensible correlation.

Dr. A. L. Boddington has clearly explained - "If all the proofs indicate that the two variables are correlated or may be correlated, even then all the proofs be carefully examined.

The change in income is the cause while increase in expenditure is the effect; increase in price is the cause, while decrease in demand is the effect; unemployment is the cause, increase in crime is the effect. Increase in vehicles is the cause while increase in accidents is the effect. Therefore, when there is a cause and effect relationship, the correlation is said to be significant.

Importance of Correlation

Correlation principle and technique are considered very important in statistics. The credit to develop and give modern coverage to this principle goes to Francis Gallon and Karl Pearson, who analysed many problems of Biology and Genetics, with the help of this technique. This is also known with the help of this principle how much and which type of relation (positive or negative) exists between two connected variables. Useful methods like regression analysis and ratio-variation principles are based upon the technique of correlation. Reliable estimation of possible value of a variable can be calculated on the basis of the definite value of the other variable with the help of these. It helps business executives and others in estimating sales, cost, demand etc. Principle of correlation is used in finding salary and cost of living index, sale and profit etc., in practical life. Therefore **Tippet** says, "The effect of correlation is to reduce the expansion of the basis of our uncertainty in prophecy." Thus predictions based on this principle is much reliable and near to reality.

11.2 Types of Correlation

On the basis of the direction of change, ratio and number of variables correlation may be of following types :

1. Positive and Negative
2. Simple, Multiple and Partial
3. Linear, Non Linear or Curvilinear

1. Positive and Negative Correlation : If the increase (decrease), in one variable is followed by the increase (decrease) in the other variable, the correlation is said to be positive or Direct. On the other hand, if the increase (decrease) in one variable causes decrease (increase) in the other variable, the correlation is said to be negative or inverse, In other words when both the variables changes in same direction, then the relationship between the two variables is called positive or direct. But when the changes are in the opposite directions that is one increasing and the other is decreasing, the correlation is negative or inverse. For determining the direction of change average values are taken.

Examples- (Positive Correlation) : -

1. Increase in rainfall (upto a point) and production of rice.
2. Higher amount of capital employed and higher expected profit.
3. Progress in business and employment.

Examples - (Negative correlation) : -

1. Demand of a commodity may go down as a result of rise in prices.
2. Increase in the number of television sets and the number of cinema goers.

2.Simple, Multiple and Partial Correlation : -

(i) **Simple correlation :** Correlation between any two data series is called simple correlation. One of these series is 'cause' or independent variable and the second series is the 'effect' or dependent variable.

(ii) **Multiple correlation :** If the common effect of two or more independent variables on dependent data series is studied, the relationship is called multiple correlation. In this there is more than one independent series whereas dependent series is only one; e.g., the study of the effect of quantity of rain, average temperature, nature of soil and capacity of labourers on wheat production per acre. In this example, figures of wheat production per acre form dependent data series; and rain, temperature, nature of soil, capacity of labourers make independent series. Therefore the correlation between these factors can be known only by multiple correlation.

(iii) **Partial correlation :** In this more than two data series are studied but by keeping the effect of other data series constant, values of only two variables are studied together. Knowing that wheat production is affected by quantity of rain, if we study the connection between nature of soil and wheat production per acre, keeping the quantity of rain constant, it will be called partial correlation.

(3) Linear and Curvilinear or Non linear Correlation : When the changes among two variables are constant in a particular ratio, such a correlation is called linear correlation. For example, increase of every Rs. 100 in income leads to savings of Rs.30, then the correlation between income and savings will be linear. If we plot these variables on the graph paper, all the points would form a straight line; that is why it is called linear correlation.

When the rate of change is not constant, that is, at one time when income increases by Rs. 100 the saving is Rs. 40 and the second time it is Rs. 30 only and the third time it is Rs. 50, the relationship found in such cases will be non-linear or curvilinear. When values of such variables are plotted on the graph paper they will form a curve. Therefore, such a correlation is known as a curvilinear correlation. This is a type of correlation is normally found in social and economic fields.

Linear and Non linear correlation may also be positive or negative - Thus : -

(i) If changes in two variables are in the same direction and in constant ratio the correlation is Linear Positive.

(ii) If changes in two variables are in the opposite direction in constant ratio, the correlation is Linear Negative.

(iii) If changes in two variables are in the same direction but not in constant ratio, the correlation is Positive Non linear.

(iv) If changes in two variables are in opposite direction and not in constant ratio, the correlation is Negative Curvilinear.

Activity - A

State whether the following data have positive / negative or linear / curvilinear correlation : -
i) Every 10% increase in inflation results in 20% increase in general price level.
ii) Every 10% increase in price of a commodity is associated with 5% decrease in demand.
iii) For every 10% increase in the quantity of money in circulation, the general price level increases by 5%, 8%, 9% etc.
iv) For every 10% increase in the price of commodity, 5 to 10% decrease in demand.

11.3 Degree of Correlation

By an inspection of the variables, we can know about the direction of correlation. But for the numerical value we have to calculate coefficient of correlation from which any significant conclusion can be drawn. Wherever the term correlation is used, it refers to coefficient of correlation. Degrees of coefficient of correlation are as follows :

1. Perfect Correlation
2. Partial Correlation
   - (i) High
   - (ii) Moderate
   - (iii) Low
3. Absence of Correlation

1. Perfect correlation : If the change of two data series takes place in the same direction and in equal proportion, there exists perfect positive correlation between them. Correlation in such a case is +1, e.g., if the amount of electricity bill increases in equal proportion definitely based on the increase of number of consumption units. On the opposite if the change of values of two data series is in equal proportion but not in same direction, there exists perfect negative correlation between them. In such a case coefficient of correlation is -1. For example, if increase in price of a commodity by 20% leads to decreases 20% of its demand, it will be perfect negative correlation.

Normally perfect positive and negative correlations are found in natural sciences while in social and economic fields we do not find such relations.

2. Limited degree of correlation : When there is neither absence of correlation nor perfect correlation between the two variables, it is called limited degree of correlation. This may be either positive or negative and the coefficient comes between zero and one. Correlation based on limits can be of three types :

(i) **High Degree Correlation :** When coefficient of correlation is between .75 to 1 then it is said to be high degree correlation. It may also be positive or negative depending upon the positive or negative values of coefficient.

(ii) **Moderate Degree Correlation :** When coefficient of correlation is in middle i.e. neither very high nor very low, that is, between .3 to .75 then it is said to be a moderate degree correlation. If the coefficient is positive, it is said to be moderate degree positive and if it is negative, it is called moderate degree negative correlation.

(iii) **Low Degree Correlation :** When coefficient of correlation is between 0 and .3 then it is called low degree correlation. Depending upon the positive or negative value of coefficient, it may also be low degree positive or low degree negative.

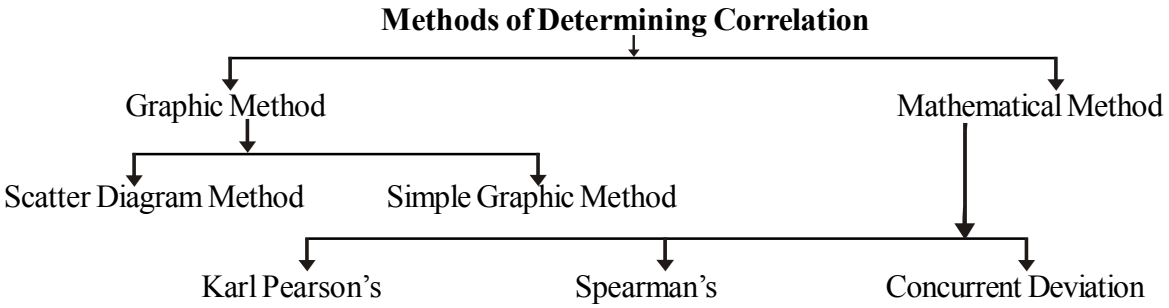
3. Absence of Correlation : When the variables do not have any type of relationship either in the same direction or in the opposite direction then it is said to be absence of correlation. If coefficient of correlation of such series is calculated, it will come to zero.

Degree of correlation in one sight

Degree	Positive	Negative
Perfect	(+1)	(-1)
Height	+0.75 to +1	-.75 to -1
Moderate	+.3 to +0.75	-0.3 to -0.75
Low	0 to +.3	0 to -.3
Absence	zero (0)	zero (0)

11.4 Methods of determining Correlation

Following methods are used for studying the coefficient of correlation : -



Graphic Methods

11.4.1 Scatter Diagram

Scatter diagram method is an elementary method of knowing the direction of the two variables. This is able to visually show the relationship between two variables. We take X variable on X axis (horizontal axis) and Y variable on Y axis (vertical axis). All pairs of X and Y are plotted on the graph paper in such a way as there is the same dot for both the values. So the number of value pairs is equal to the number of dots on graph paper.

For example we may have figures on advertisement expenditure (X) and sales (Y) of a firm for the last 8

years. When this data is plotted on a graph, we obtain a scatter diagram. By looking to the scatter of different points one can form an idea about the fact whether the variables are related or not. The greater the scatter of the plotted points on the graph paper, the lesser is the relationship between the two variables. If the dots marked on graph paper make a simple and straight line from lower left hand corner to upper right hand corner, correlation is perfect positive or (+1) on the other hand if all the points are lying on straight line falls from left upper corner to right lower corner, correlation would be perfect negative or (-1).

If no direction is indicated by dots in scattered diagram and they are scattered here and there, it should be taken as absence of correlation. A scatter diagram gives two very useful types of information. First, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get an idea of what kind of relationship (linear or non linear) would exists there.

Scatter Diagram of Different Degree of Correlation

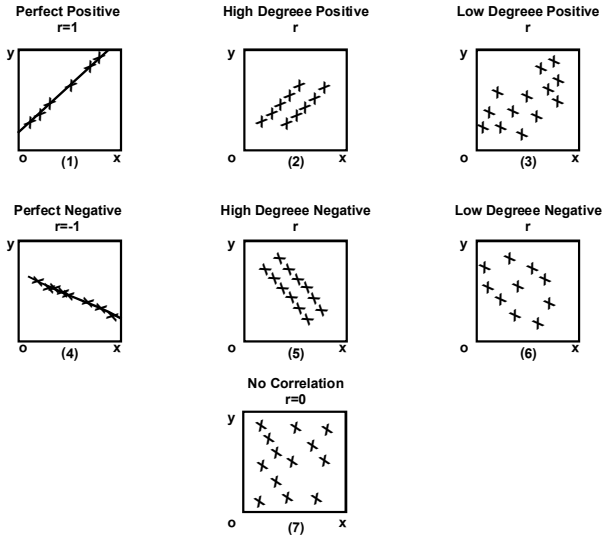


Illustration : 1

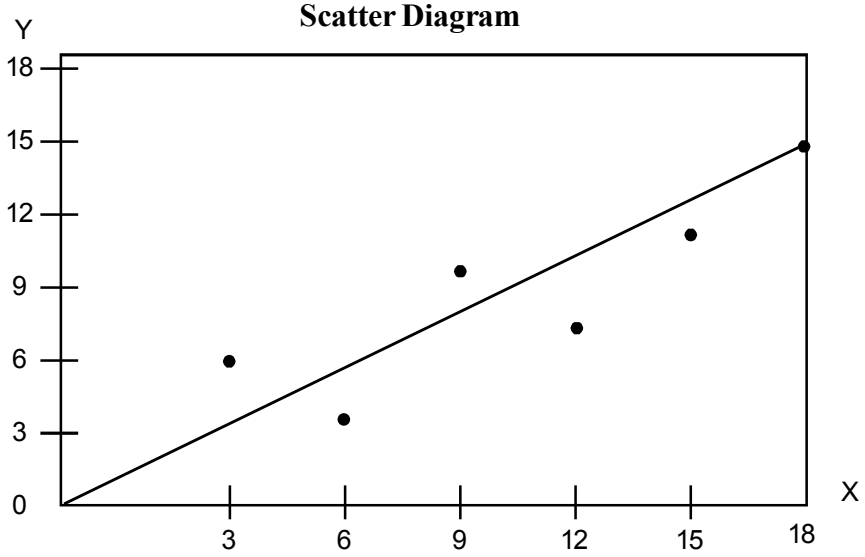
Represent the following values by Scatter Diagram and comment, whether there is any correlation between A and B : -

A	3	6	9	12	15	18
B	6	4	10	7	11	15

Solution :

Scatter diagram has been constructed by taking A on x axis (horizontal) and B on y axis (vertical).

Scatter Diagram



By inspection of the plotted points, it is clear that they are from lower left hand to upper right hand side and are not scattered. So the two variables have high degree positive correlation.

Merits : -

1. It is the first step for investigating about the relationship between two variables.

2. It is a simple and non mathematical method of finding out relationship between the two variables.

3. It is not affected by the extreme values of the series.

4. It is an additional method of verifying the conclusions drawn by mathematical approaches.

Limitations : -

The exact degree of relationship cannot be obtained by using this method. It can only reveal the direction of relationship and also whether it is high or low.

11.4.2 Simple Graphic Method

This is the second method of studying the correlation. Common factor is taken on x axis and other variables on y axis while constructing the graph. We can take one scale if both the variables are of uniform unit and if they differ, then different scales will be taken for the two variables. For this purpose, we will adjust the scales so that both the lines may be shown together.

1. If both the lines move upward from lower left hand corner to upper right hand corner then the tendency of correlation will be positive. The lesser the difference between two lines, the greater the degree of correlation.

2. If both the lines move in opposive direction, that is, one moves from lower left to right upward and then other from upper left hand to lower right hand or vice versa then the correlation will be negative. If the proportion of increase and decrease is uniform, we will find a very high degree of negative correlation.

3. If these two lines do not show any direction then this will be the case of no correlation or absence of correlation.

This method is generally used in case of time series, where the values are given for a period of time. In this method also numerical value of correlation is not obtained to study the degree of relationship.

Illustration : 2

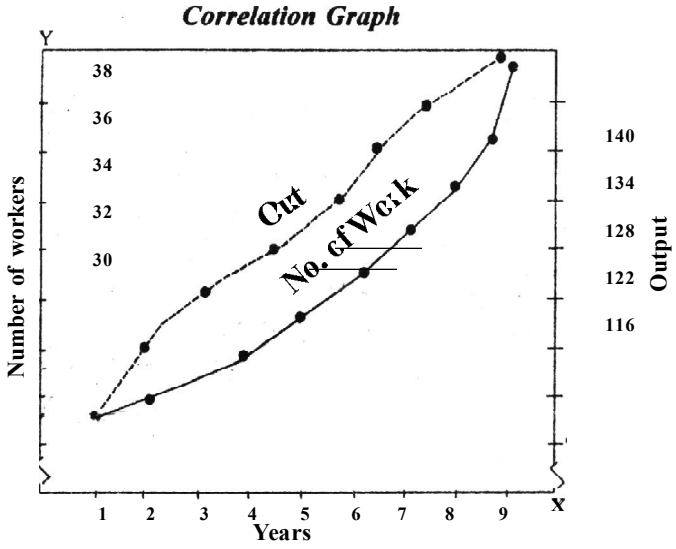
Show the following data by means of a correlation graph and comment upon the relationship of number of workers and units produced.

Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
No. of Workers	**23**	**24**	**25**	**26**	**28**	**29**	**30**	**33**	**37**
Output in'000 (Units)	**93**	**105**	**108**	**113**	**117**	**125**	**128**	**134**	**140**

Solutions : -

Years have been shown on x axis while number of workers and output have been shown on y axis. Different scales have been taken because the variables are different in units.

It is clear by having a look at the lines of the correlation graph that both curves fluctuate together in the same direction and the difference between the two lines is also not significant, therefore there exists a positive correlation.



Mathematical Methods

11.4.3 Karl Pearson's Coefficient of Correlation

Graphic method of correlation tell us about the direction and degree of correlation, but it does not measure the degree of relationship numerically. This method gives a precise and a quantitative value which can be interpreted meaningfully. The coefficient of correlation alongwith other information help in estimating the value of dependent variable from the known value of an independent variable.

Karl Pearson's method is most popular among other methods used for calculating the coefficient of correlation. This method is also known as 'Pearson's Coefficient of Correlation'. Pearson's coefficient of correlation is between $\pm$ 1. If it is +1 then the correlation is perfect positive and in case of -1, it is perfect negative. The near the value of coefficient to unity, the greater is the degree of correlation.

Computation of Karl Pearson's coefficient of correlation :

Co-variance of data series is measured before computing Karl Pearson's coefficient of correlation. This measure is called absolute degree of correlation. For relative degree of coefficient, absolute degree is divided by standard deviations of both variables and the resultant itself is called Karl Pearson's coefficient of correlation. The following formula is used : -

i) Co-variance of x and y = $\frac{\Sigma xy}{N}$

ii) Coefficient of correlation = $\frac{\text{Co-variance of x and y}}{\sigma_x \times \sigma_y}$

Calculation of Coefficient of Correlation in Individual series

Coefficient of correlation may be calculated by direct method or by short-cut method.

(1) Direct Method : Under this method, deviations are taken by actual mean and not by assumed mean. Following formula is used :

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

Simplified form of basic formula = $r = \frac{\Sigma xy}{N \sqrt{\frac{\Sigma x^2}{N}} \sqrt{\frac{\Sigma y^2}{N}}}$ **OR** $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$

Here,

r = Karl Pearson's coefficient of correlation
Σxy = Sum of the products of respective deviations of X and Y from their means.
σ_x, σ_y = Standard deviations of X and Y series respectively.
N = Number of pairs observed.

Procedure of Calculation : -

1. Calculate arithmetic mean of both the series.
2. Find out deviations from their respective means and give them $x\left(X-\overline{X}\right)$ and $y\left(Y-\overline{Y}\right)$ notations.
3. Calculate the squares of deviations and find their totals (Σx^2 and Σy^2).
4. After that product of deviations of both series and sum of this product should be obtained (Σxy).
5. Find out the standard deviations separately and apply the formula for calculating the coefficient of correlation.

Illustration : 3

Calculate Karl Pearson's Coefficient of correlation from the following data :

Ages of Husbands (years)	23	27	28	28	29	30	31	33	35	36
Ages of Wives (years)	18	20	22	27	21	29	27	29	28	29

Computation of Coefficient of Correlation

X	**x** $(x - \bar{x})$ $(\bar{x} = 30)$	**x^2**	**Y**	**y** $(y - \bar{y})$ $(\bar{y} = 25)$	**y^2**	**xy**
23	-7	49	18	-7	49	+49
27	-3	9	20	-5	25	+15
28	-2	4	22	-3	9	+6
28	-2	4	27	+2	4	-4
29	-1	1	21	-4	16	+4
30	0	0	29	+4	16	0
31	+1	1	27	+2	4	+2
33	+3	9	29	+4	16	+12
35	+5	25	28	+3	9	+15
36	+6	36	29	+4	16	+24
(Σx) **300**	**0**	(Σx^2) **= 138**	ΣY **= 250**		(Σy^2) **= 164**	Σxy **= 123**

$$\overline{X} = \frac{\Sigma X}{N} \text{ or } \frac{300}{10} \text{ or } 30; \quad \overline{Y} = \frac{\Sigma Y}{N} \text{ or } \frac{250}{10} \text{ or } 25$$

Standard Deviation of X series

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} \text{ or } \sqrt{\frac{138}{10}} = 3.7148$$

Standard Deviation of Y series

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} \text{ or } \sqrt{\frac{164}{10}} = 4.0497$$

By substituting the values in the original formula,

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

$$r = \frac{123}{10 \times 3.71 \times 4.05} = \frac{123}{150.255} = +0.8186$$

So there is very high degree of positive correlation between the age of husbands and age of wives.

Alternate Formula :

In the above illustration, the coefficient of correlation may be calculated by changed formula in the following manner : -

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

$$r = \frac{123}{\sqrt{138 \times 164}} = \frac{123}{\sqrt{22632}}$$

$$= \frac{123}{150.44} = +0.8186$$

Both the formulae give same result. But for easy calculations formula II is more popular in practice.

Illustrations : 4

i) Coefficient of correlation between two variables X and Y is 0.28, their co-variance is 7.6 and the variance of X is 9, find the standard deviation of Y series.

ii) If the co-variance between X and Y variables is 1352 and the variance of X and Y are 2699 and 1112 respectively, find r between the two variables.

iii) Find out the number of items if :

$r = .5$, $\Sigma xy = 60$, $\sigma_y = 4$ and $\Sigma x^2 = 90$

Solution :

i) Given ; r = .28, Co-variance $= \frac{\Sigma xy}{N} = 7.6$

Variance of X = 9, $\sigma_y = ?$

$$r = \frac{\Sigma xy}{N \times \sigma_x \times \sigma_y} \text{ or } .28 = \frac{7.6}{\sqrt{9} \times \sigma_y} = .28 \times 3 \times \sigma_y = 7.6$$

$$\text{or } \sigma_y = \frac{7.6}{3 \times .28} = 9.05$$

ii) $$r = \frac{\text{co-variance of X and Y}}{\sqrt{(\text{Variance of X})} \times \sqrt{(\text{Variance of Y})}} = \frac{1352}{\sqrt{2699 \times 1112}}$$

$$= \frac{1352}{1732.42} = +0.7804$$

iii) Given : r = .5, $\Sigma xy = 60$, $\sigma_y = 4$ and $\Sigma x^2 = 90$

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

$$.5 = \frac{60}{N \times \sqrt{\frac{90}{N}} \times \sqrt{16}} \quad \text{by squaring we get}$$

$$.25 = \frac{3600}{N \times N \times \frac{90}{N} \times 16} = \frac{3600}{1440N}$$

.25 x 1440N = 3600

$$N = \frac{3600}{25 \times 1440} = 10$$

Computation of correlation by squares of values : -

Calculation of coefficient of correlation is possible even without obtaining deviations from assumed mean or actual mean in the following manner : -

i) Find totals of each series (ΣX and ΣY).

ii) Square of each value in both the series and find total (ΣX^2 and ΣY^2).

iii) Product of respective values of X and Y and total (ΣXY).

Formula : -

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\ \sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

This method is useful when values are too small.

Use of Logarithms - for the purpose of multiplication, square root and divisions, logs are used to make the calculations easy.

In illustration-3, the coefficient of correlation may be calculated by using log values in the following manner:

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

r = Antilog [log of Σxy - ½ (log of Σx^2 + log of Σy^2)]
r = Antilog [log of 123 - ½ (log of 138 + log of 164)]
r = Antilog [2.0899 - ½ (2.1399 + 2.2148)]
r = Antilog [2.0899 - ½ (4.3547)]
r = Antilog (2.0899 - 2.17735)
r = Antilog $\bar{1}.91165$
r = +.81

(2) Short-cut method

In direct method deviations are taken from real means. If the mean is not a whole number but in decimal, it will be difficult to compute deviations and their squares. Therefore short cut method should be used.

Calculation of r with the help of assumed mean is as follows : -

First Formula

$$r = \frac{\sum dxdy - N\left(\overline{X} - A_1\right)\left(\overline{Y} - A_2\right)}{N \times \sigma_x \times \sigma_y}$$

Here :

dx = Deviation of X series from its assumed mean
dy = Deviations of Y series from its assumed mean
Σdx^2 & Σdy^2 = Sum of deviations got from assumed mean of two variables
$\Sigma dxdy$ = Sum of product of deviations
N = Number of pairs
$\overline{X}$ *and* $\overline{Y}$ = Arithmetic mean of X and Y series
A_1 and A_2 = Assumed mean of X and Y series

Procedure of Calculation

1. Calculate deviations (dx and dy) of X and Y series by assuming an appropriate and convenient number.

2. Find the squares of these deviations and calculate standard deviations.

3. Multiply the deviations and find the sum of product ($\Sigma dxdy$)

4. Calculate actual mean with the help of Σdx and Σdy. Find the sum of all the values (dx, dy, dx^2, dy^2 and dxdy) and then use the above formula to calculate coefficient of correlation.

5. If one of the variable or size is in groups then, mid value is used as x or y respectively.

Second Formula

$$r = \frac{\sum dxdy - \left(\frac{\sum dx \times \sum dy}{N}\right)}{\sqrt{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}} \times \sqrt{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}}}$$

In this formula we are not required to calculate the arithmetic mean separately, but division process is too much so it is not popular in practice.

Third Formula

$$r = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sqrt{\sum dx^2 \times N - \left(\sum dx\right)^2} \times \sqrt{\sum dy^2 \times N - \left(\sum dy\right)^2}}$$

Illustration : 5

The following are the results of B.Com. examination :

Age of Candidates	% of Failures	Age of Candidates	% of Failures
17	38	22	37
18	40	23	42
19	35	24	46
20	32	25	52
21	34	26	56

Calculate coefficient of correlation between age and success in the examination.

Solution :

Correlation has been calculated by first finding out the percentage of successful boys which is 100% of failures.

Calculation of Coefficient of Correlation

Age of Candidates	A = 21 dx	dx^2	% of Successful Candidates Y	A = 60 dy	dy^2	dxdy
17	-4	16	62	+2	4	-8
18	-3	9	60	0	0	0
19	-2	4	65	+5	25	-10
20	-1	1	68	+8	64	-8
21	0	0	66	+6	36	0
22	+1	1	63	+3	9	+3
23	+2	4	58	-2	4	-4
24	+3	9	54	-6	36	-18
25	+4	16	48	-12	144	-48
26	+5	25	44	-16	256	-80
	+5	**85**		**-12**	**578**	**-173**

By applying Karl Pearson's formula :

$$r = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sqrt{\sum dx^2 \times N - \left(\sum dx\right)^2} \times \sqrt{\sum dy^2 \times N - \left(\sum dy\right)^2}}$$

$$r = \frac{-173 \times 10 - \left(5 \times -12\right)}{\sqrt{85 \times 10 - \left(5\right)^2}\sqrt{578 \times 10 - \left(-12\right)^2}} = \frac{-1730 + 60}{\sqrt{850 - 25}\sqrt{5780 - 144}}$$

$$r = \frac{-1670}{\sqrt{825}\sqrt{5636}} = \frac{-1670}{28.72 \times 75.07} = \frac{-1670}{2156}$$

r = -.7746

Illustration : 6

On the basis of following information, find out if there is any relation between age and illiteracy.

Age Group	Total Population (in'000)	Illiterate Population (in'000)
10-20	240	200
20-30	200	150
30-40	160	120
40-50	100	60
50-60	50	40
60-70	30	20
70-80	10	10

Solution :

Calculation of Coefficient of Correlation

Age Group	M.V. X	dx A = 45	Step i = 10	dx^2	Illiterates Popu. '000	dy A = 80	dy^2	dx × dy
10-20	15	-30	-3	9	83	+3	9	-9
20-30	25	-20	-2	4	75	-5	25	+10
30-40	35	-10	-1	1	75	-5	25	+5
40-50	45	0	0	0	60	-20	400	0
50-60	55	+10	+1	1	80	0	0	0
60-70	65	+20	+2	4	67	-13	169	-26
70-80	75	+30	+3	9	100	+20	400	+60
Total			**0**	**28**		**-20**	**1028**	**+40**

Illiterate Population per 1000 = $\frac{Illiterate\ Population}{Total\ Population} \times 1000$

$$r = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sqrt{\left[\sum dx^2 \times N - \left(\sum dx\right)^2\right]\left[\sum dy^2 \times N - \left(\sum dy\right)^2\right]}}$$

$$r = \frac{40 \times 7 - \left(0 \times -20\right)}{\sqrt{\left[28 \times 7 - \left(0\right)^2\right]\left[1028 \times 7 - \left(20\right)^2\right]}} = \frac{280 - 0}{\sqrt{(196 - 0)(7196 - 400)}}$$

$$r = \frac{280}{\sqrt{196 \times 6796}} = \frac{280}{\sqrt{1332016}} = \frac{280}{1154.13} = 0.2426$$

Correlation in Discrete or Grouped Series or in Bi-variate Frequency Series

Like individual observations the coefficient of correlations can also be calculated in discrete or in grouped series. For this purpose bi-variate correlation table is to be formed in which cell frequencies of discrete or continuous series are present in such a way as to classify their relation.

First Formula (Direct method)

$$r = \frac{\sum fxy}{N \times \sigma_x \times \sigma_y}$$

Here,

Σfxy = Summation of the cell frequencies multiplied with corresponding deviations of X and Y from actual means.

N = Sum of all the cell frequencies or Σf.

Generally the value of mean is in fraction, so short-cut method is used in place of direct method. As 'f' has been multiplied the above formula, in the same way 'f' can be multiplied in all the short-cut method formula discussed earlier.

Second Formula (short cut method)

$$r = \frac{\sum fdxdy - \frac{\left(\sum fdx\right) \times \left(\sum fdy\right)}{N}}{\sqrt{\sum fdx^2 - \frac{\left(\sum fdx\right)^2}{N}} \times \sqrt{\sum fdy^2 - \frac{\left(\sum fdy\right)^2}{N}}}$$

Third Formula (Short-cut method)

$$r = \frac{\sum fdxdy \times N - \left(\sum fdx \times \sum fdy\right)}{\sqrt{\sum fdx^2 \times N - \left(\sum fdx\right)^2} \times \sqrt{\sum fdy^2 \times N - \left(\sum fdy\right)^2}}$$

Procedure of Calculation : -

1. Construct a correlation table in which four columns at downward and four rows at the right hand side for f, fdy, fdy^2, fdxdy (four columns); and f, fdx, fdx^2 and fdxdy (four rows) respectively.

2. dx and dy are calculated by taking a convenient value or midpoint as assumed mean separately for X and Y series. If one of the series is discrete then at that side we will take one column for deviations only. If class interval is equal, data deviation can be calculated by a common factor. In case of unequal class intervals of two series also (e.g., class range in series X is 5 and in series Y is 10) step deviation can be calculated. Deviations can be taken only in one series also.

3. Find out fdy and fdx by multiplying deviations with the frequency of the respective group.

4. Find fdy^2 and fdx^2 by multiplying fdy and fdx with respective deviations.

5. For calculating fdxdy, the respective deviations (dx and dy) are multiplied and the product is written on the left hand upper side of every cell. Thus multiplied, figure of dxdy is again to be multiplied with the frequency of that particular cell and written in right hand lower corner; it will be the figure of fdxdy.

6. Now find the sum of all the columns and rows as N or (Σf), Σfdy, Σfdy^2, Σfdxdy, Σfdx, Σfdx^2 and Σfdxdy and substitute values in the formula.

Illustration :7

In a survey of 100 school teachers of a city following data were obtained regarding their income and savings. Calculate correlation between income and savings.

Income (Rs.)	Saving (Rs.)				Total
	150	200	250	300	
1000	8	4	-	-	12
1400	-	12	24	6	42
1800	-	9	7	2	18
2200	-	-	10	5	15
2600	-	-	9	4	13
Total	**8**	**25**	**50**	**17**	**100**

Solution :

Computation of Coefficient of Correlation

$$r = \frac{\sum fdxdy \times N - (\sum fdx \times \sum fdy)}{\sqrt{\sum fdx^2 \times N - (\sum fdx)^2}\ \sqrt{\sum fdy^2 \times N - (\sum fdy)^2}}$$

$$r = \frac{34 \times 100 - (-25 \times 76)}{\sqrt{157 \times 100 - (-25)^2}\ \sqrt{126 \times 100 - (76)^2}}$$

$$r = \frac{3400 + 1900}{\sqrt{15700 - 625}\ \sqrt{12600 - 5776}} = \frac{5300}{\sqrt{15075 \times 6824}}$$

$$r = \frac{5300}{122.78 \times 82.608} = \frac{5300}{10142.6}$$

$$r = +0.5225$$

Income (Rs.) X	Saving (Rs.) Y		150	200	250	300				
	dx (A= 1800)	dy (A=200)	-50	0	+50	+100				
		Step i = 50 i = 400	-1	0	+1	+2	Total	fdx	fdx^2	fdx . dy
1000	-800	-2	+2 8 +16	0 4 0	–	–	12	-24	48	+16
1400	-400	-1	–	0 12 0	-1 24 -24	-2 6 -12	42	-42	42	-36
1800	0	0	–	0 9 0	0 7 0	0 2 0	18	0	0	0
2200	+400	+1	–	–	1 10 +10	2 5 +10	15	15	15	+20
2600	+800	+2	–	–	2 9 +18	4 4 +16	13	26	52	+34
		Total	8	25	50	17	100	-25	157	+34
		fdy	-8	0	50	34	76			
		fdy^2	8	0	50	68	126			
		fdx . dy	16	0	+4	+14	+34			

Illustration : 8

Calculate the coefficient of correlation between ages of husbands and wives.

Ages of Husbands (Y)	Ages of Wives (X)					Total
	20-30	30-40	40-50	50-60	60-70	
25-35	6	3	-	-	-	9
35-45	3	16	10	-	-	29
45-55	-	10	15	7	-	32
55-65	-	-	7	10	4	21
65-75	-	-	-	4	5	9
Total	**9**	**29**	**32**	**21**	**9**	**100**

Solution :

Computation of Coefficient of Correlation

$$r = \frac{\sum fdxdy \times N - \left(\sum fdx - \sum fdy\right)}{\left[\sum fdx^2 \times N - \left(\sum fdx^2\right)\right]\left[\sum fdy^2 \times N - \left(\sum fdy^2\right)\right]}$$

$$r = \frac{98 \times 100 - \left(-8 \times -8\right)}{\sqrt{\left[100 \times 122 - \left(-8\right)^2\right]\left[100 \times 122 - \left(-8\right)^2\right]}}$$

$$r = \frac{9800 - 64}{\sqrt{\left[12200 - 64\right]\left[12200 - 64\right]}}$$

$$r = \frac{9736}{\sqrt{12136 \times 12136}} = +0.802$$

Husband's Age (Y)	Ages of Wives (X)							f	fdy	fd^2y	fdxdy
	Mid Value		20-30	30-40	40-50	50-60	60-70				
			25	35	45	55	65				
		dy \ dx	-2	-1	0	+1	+2				
25 - 35	30	-2	4 6 24	2 3 6	–	–	–	9	-18	36	30
35 - 45	40	-1	2 3 +6	1 16 +16	0 10 0	–	–	29	-29	29	22
45 - 55	50	0	–	0 10 0	0 15 0	0 7 0	–	32	–	–	–
55 - 65	60	+1	–	–	0 7 0	1 10 10	2 4 8	21	21	21	18
65 - 75	70	+2	–	–	–	2 4 8	4 5 20	9	18	36	28
	F		9	29	32	21	9	100 (N)	(-) 8 Σfdy	122 Σfd^2y	98 Σfdxdy
	fdx		-18	-29	0	21	18	(-) 8 Σfdx			
	fd^2x		36	29	0	21	36	122 Σfdx^2			
	fdxdy		30	22	0	18	28	98 Σfdxdy			

Assumptions of Karl Pearson's Coefficient of Correlation

Pearson's coefficient of correlation is based on the following assumptions :

1. Linear Relationship : - Both the variables have linear relationship. If we plot these variables on scatter diagram we will find a straight line.

2. Casual Relationship : There is a cause and effect relationship between the two variables. If there is a lack of cause and effect relationship and even then there is correlation between the two variables, such a correlation is said to be spurious or nonsense correlation.

3. Normality : Both the variables form a normal distribution because these are affected by a number of independent causes.

Properties of the Coefficient of Correlation

Following are the important properties of coefficient of correlation : -

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

2. The coefficient of correlation is independent of change of scale and origin of the variables X and Y. Change of scale means taking deviations by any value from the given values of X and Y and change of origin means taking step deviations by any common value, i.e., by dividing or multiplying the values by any constant value.

3. The coefficient of correlation is the geometric mean of two regression coefficients.

$$r = \sqrt{b_{xy} \times b_{yx}} \ .$$

b_{xy} and b_{yx} are regression coefficients of X on Y and Y on X respectively (for details, see next unit)

Merits of Karl Pearson's Method

Karl Pearson's method is the best method among the mathematical methods used for calculating the coefficient of correlation. This method measures not only the degree of correlation but also the direction of correlation.

This method along with other information helps in obtaining the estimated value of dependent variables from the known value of independent variable.

Demerits of Karl Pearson's Method

1. The coefficient of correlation assumes linear relationship between the two variables, which is not possible in every relation.

2. If the data are not reasonably homogenous, then it may give misleading picture of the degree of relationship.

3. The value of coefficient is unduly affected by the extreme values.

4. It takes more time to calculate the value of correlation as compared to other method.

11.4.4 Probable Error and Standard Error

In modern time most of the statistical investigations are done by sampling method. Therefore it is quite natural to get the difference between the two samples taken from one universe. But the main question is whether this difference is significant or not significant. If the difference is insignificant, it may be ignored. Correlation calculated from one sample may differ significantly from the correlation of other sample. To determine this fact probable error is calculated.

Probable error is that measurement or amount by adding and subtracting of which in the coefficient of correlation two such limits are derived as there is equal chance of finding coefficient of correlation of the random samples or whole data of the same nature.

According to **Wheldon,** "Probable error defines the limit above and below the size of the coefficient determined within which there is an equal chance that coefficient of correlation, similarly calculated from other samples, will fall."

In the words of **Horace Secrist**, "The probable error of r is a value which if added to and subtracted from the average correlation coefficient, produces limits within which the chances are even that a coefficient of correlation from a series selected at random will fall."

The probable error of the coefficient of correlation helps in interpreting its value. The reliability of the value of coefficient can be determined with its help so far as it depends on the conditions of random sampling.

The probable error of r is obtained with the help of the following formula.

$$\text{Probable Error or P.E.} = .6745\frac{1-r^2}{\sqrt{N}}$$

Illustration : 9

If the coefficient of correlation between two variables is +0.8 and the number of items are 25, find out probable error.

Solution :

$$P.E. = .6745\frac{1-r^2}{\sqrt{N}} = .6745\frac{1-(.8)^2}{\sqrt{25}}$$

$$.6745\frac{1-.64}{5} = \frac{.6745 \times .36}{5} = \frac{.1282}{5} = .0485$$

Utility of Probable Error

Following are two main uses of probable error:

1. Determination of Limits of r : Probable error determines the two limits of coefficient of correlation with in which the coefficient of correlation of other samples or coefficient of correlation of the universe is expected to lie. For calculating these limits, the probable error is added and subtracted from coefficient of correlation. It may be written as follows:

Coefficient of correlation ± Probable Error, i.e., r ± P.E.

2. Interpretation of Coefficient of Correlation: Probable error is also used for interpreting the significance of coefficient of correlation. The following five rules will decide whether coefficient of correlation of any data series is important or not :

(i) If coefficient of correlation (r) is less than P.E. (r < P.E.), it proves that there is lack of correlation between the two series.

(ii) If coefficient of correlation (r) is more than six times of P.E. (r > 6 P.E.), the correlation between the data series exists definite and significant.

(iii) If the coefficient of correlation is more than 3 times but less than six times of P.E., the correlation will exist but it will not be considered significant.

(iv) If the coefficient of correlation is less than 3 times of P.E., existence of correlation will be doubtful.

(v) If the coefficient of correlation is less than 0.3 and P.E. is too little, correlation between the series will exist but it will not be significant or important.

Limitations of Probable Error

The use of probable error is suitable in the following cases : -

1. The sample must be unbiased.
2. The sample must be taken from a universe.
3. The data must follow the conditions of a normal curve.
4. The number of pairs of items must be adequate.

In Social, Economical and Business areas these conditions are not satisfied hence the significance of coefficient of correlation cannot be judged properly by probable error. Moreover, the test of existence of correlation on the basis of probable error is true only to the extent of 50%.

Standard Error

Since probable error determines the limits of probable 50% cases of coefficient of correlation and its results are also not good in socio-economic and business fields. Hence, now-a-days standard error is used in place of probable error.

The following is the formula : -

$$\text{Standard Error or S.E.} = \frac{1-r^2}{\sqrt{N}} \; or \; P.E. \times \frac{3}{2}$$

For determining the limits of coefficient of correlation of a sample in order to cover almost all the cases of universe the following formula is used; r ± 3 S.E.of r.

Illustration : 10

By taking 625 samples out of total husbands and wives, correlation of weight between them is 0.5. Find possible limits of coefficient of correlation of the whole.

Solution :

Limits of coefficient of correlation = r ± 3 S.E. of r

$$\text{S.E. of r} = \frac{1-r^2}{\sqrt{N}} = or \; \frac{1-(.5)^2}{\sqrt{625}} = \frac{0.75}{25} = 0.03$$

Limits : r + 3 S.E. = 0.5 + 3 x 0.03 = 0.59

r - 3 S.E. = 0.5 - 3 x 0.03 = 0.41

Therefore coefficient of correlation will lie between 0.59 and 0.41.

11.4.5 Spearman's Ranking Method

Prof. Charles Spearman developed this method of finding out correlation between two individual or ungrouped variables in 1904. This method is also known as Spearman's Rank Differences Method or Ranking Method.

Sometimes we cannot measure certain facts quantitatively but can put them in some definite rank, e.g., wisdom, honesty, health, beauty, ability, character, skill etc. Where it is not possible to measure in perfect quantities due to absence of numerical form or when the shape of distribution is not known ranking figures are used there. These ranks are determined according to the size of data. This method is much easier than Karl Pearson's method. For example, it is very difficult to depict beauty in the form of numbers but things or persons can be arranged according to rank in beauty. If this ranking is done by two or more persons about the same group, differences are often found because there is no definite scale of evaluation. It is affected by the personal point of view of the judges. Coefficient of correlation is find out on the basis of these ranks by ranking method. The value of rank coefficient is also between -1 and +1.

Rank correlation may be studied in two situations:

1. When ranks are not given.
2. When ranks are given.

When Ranks are Not Given

The following method is used for calculating the coefficient of correlation :

1. Assigning Ranks : - The ranks are assigned for all values of X as well as Y variables. While awarding the ranks, the first rank is given to the maximum value and the second rank to the next lower value and in the same way the last rank to the lowest value. Ranks may also be started from the lowest value to the highest value. But the same method should be followed for both the variables.

2. Calculating Rank Differences : - For calculating the rank differences, the ranks assigned to Y variable (R_2) will be subtracted from the respective ranks of X (R_1) and Rank difference are taken as R_1-R_2 = D. The sum of rank differences is always equal to zero.

3. These differences will be squared and the sum of these squares will be ΣD^2.

4. Following formula will be used for calculating the rank correlation :

$$r_R = 1 - \frac{6\sum D^2}{N\left(N^2 - 1\right)} \text{ or } 1 - \frac{6\sum D^2}{\left(N^3 - N\right)}$$

Here, r_R = Coefficient of rank correlation

ΣD^2 = Sum of squares of differences between ranks

N = Number of pairs of observations.

5. Correction for Tied Ranks : In case of repeated or common values, the procedure for allotment of ranks is slightly different. In such circumstances common ranks are given to the repeated values. The common rank is the average of ranks which these items would have assigned if they were slightly different from each other. Next item will get the rank next to the rank already assumed. In this situation the following adjustment or correction in the above formula is applied :

$$\frac{1}{12}\left(m^3 - m\right)$$

Here, m is the number of ranks repeated. The above factor is to be added for each repeated value. Thus formula for the calculation will be as follows : -

$$r_R = 1 - \frac{6\left[\left(\sum d^2\right) + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right).....\right]}{N\left(N^2 - 1\right)}$$

Illustration : 11

Calculate the Spearman's coefficient of rank correlation from the following data :

X	57	38	45	12	20	20	70	31	20	62
Y	12	12	30	6	14	4	24	10	6	18

Solution :

Calculation of Spearman's Coefficient of Correlation

X	Y	R_1	R_2	D	D^2
57	12	3	5.5	-2.5	6.25
38	12	5	5.5	-0.5	0.25
45	30	4	1	+3.0	9.00
12	6	10	8.5	+1.5	2.25
20	14	8	4	+4.0	16.00
20	4	8	10	-2.0	4.00
70	24	1	2	-1.0	1.00
31	10	6	7	-1.0	1.00
20	6	8	8.5	-0.5	0.25
62	18	2	3	-1.0	1.00
N = 10				Σ D = 0	ΣD^2 = 41.00

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

$$= 1 - \frac{6\left[41 + \frac{1}{12}\left(3^3 - 3\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\right]}{10^3 - 10}$$

$$= 1 - \frac{6\ (41+2+.5+.5)}{990} = 1 - \frac{6 \times 44}{990}$$

$$= 1 - \frac{264}{990} \text{ or } \frac{726}{990} = .733$$

When Ranks are Given

When ranks are already given, we have not to assign ranks. The rest of the process is the same. In some cases, we may have more than two variables but the correlation will be calculated by taking two variables at a time, we may have different combinations in such cases. It will be clear from the following example.

Illustration : 12

Ten competitors in a beauty contest are ranked by three judges in the following order. Use the rank correlation coefficient to discuss which pair of judges has the nearest approach to common taste in beauty.

First Judge	:	1	6	5	10	3	2	4	9	7	8
Second Judge	:	3	5	8	4	7	10	2	1	6	9
Third Judge	:	6	4	9	8	1	2	3	10	5	7

Solution :

Rank correlation between 1 and 2 judge

$$r_R = 1 - \frac{6 \sum D^2}{N\left(N^2 - 1\right)} = 1 - \frac{6 \times 200}{10\ \left(10^2 - 1\right)} = 1 - \frac{1200}{990} = - \frac{210}{990} = - .212$$

Rank correlation between 2 and 3 judge

$$r_R = 1 - \frac{6 \sum D^2}{N\ \left(N^2 - 1\right)} = 1 - \frac{6 \times 214}{10\ \left(10^2 - 1\right)} = 1 - \frac{1284}{990} = - \frac{294}{990} = - .2969$$

Rank correlation between 1 and 3 judge

$$r_R = 1 - \frac{6 \sum D^2}{N\ \left(N^2 - 1\right)} = 1 - \frac{6 \times 60}{10\ \left(10^2 - 1\right)} = 1 - \frac{360}{990} = \frac{630}{990} = + .6363$$

Pair of 1st and 3rd judges has the nearest approach to common taste of beauty.

Calculation of Coefficient of Rank Correlation

R_1	R_2	R_3	Deviation Between 1 & 2 Ranks		Deviation Between 2 & 3 Ranks		Deviation Between 1 & 3 Ranks	
			D_1	D_1^2	D_2	D_2^2	D_3	D_3^2
1	3	6	-2	4	-3	9	-5	25
6	5	4	+1	1	+1	1	+2	4
5	8	9	-3	9	-1	1	-4	16
10	4	8	+6	36	-4	16	+2	4
3	7	1	-4	16	+6	36	+2	4
2	10	2	-8	64	+8	64	0	0
4	2	3	+2	4	-1	1	+1	1
9	1	10	+8	64	-9	81	-1	1
7	6	5	+1	1	+1	1	+2	4
8	9	7	-1	1	+2	4	+1	1
Total			**0**	**200**	**0**	**214**	**0**	**60**

Merits of Rank Difference Method :

i) Computation of coefficient of correlation by rank method is very easy.

ii) If actual values of a series are unknown but ranks are known, correlation coefficient can be calculated.

iii) This method is much appropriate, when importance is given only to difference of ranks of individual data.

iv) This method is specially useful in calculation of correlation coefficient in qualitative facts and irregular data series.

v) This method can be used to study the degree of association between two attributes, where individuals can be ranked in some order without difficulty.

vi) This method is more useful in such cases where number of items to be studied are less.

Demerits :

i) This method is not useful in case of discrete or grouped series.

ii) It becomes difficult to use this method if number of data in a series is too large or same ranking numbers are too many.

iii) In this method absolute values have no importance.

iv) Results obtained by this method is approximate because original values are not considered.

11.4.6 Concurrent Deviation Method

Concurrent deviation method is much useful to know correlation of short-time changes in time series. Sometimes there is no need to calculate actual degree of correlation between two data series but it is much important to know the direction of correlation, ie., positive or negative. For this purpose concurrent deviation method is appropriate.

The basis of this method also is like a graph. When two variables move in the same direction on a graph, they show a positive correlation. In the same way, if the deviations of X and Y variables are in the same direction, the correlation will be positive and contrary to this, it will be negative. Following are the characteristics of this method in comparison to other methods :

1. Under this method deviations are derived by comparison of previous value and not from actual or assumed mean.

2. Only the directions of deviations (positive or negative) are considered and not the actual values or degree of deviations.

3. Under this method short term changes are paid attention and not trend.

Procedure of Calculation

(i) In this method each value of a series is compared to its previous one. If the number is bigger than the previous number, the deviation will be positive (+); and if less than that, the deviation will be negative (-); and if equal to that, there will be no deviation and so sign of equality (=) will be used. For first value, there will be no deviation sign.

ii) After giving signs, we shall find concurrent signs from both the variables, for this purpose variables having the same signs. i.e., either + or - or =, shall be given + sign.

iii) After that deviation of same direction, i.e. signs of (+) are added and their total number is calculated. This very number is called concurrent deviation number, and then the following formula is used : -

$$r_c = \pm\sqrt{\pm\frac{(2c - N)}{N}}$$

Here,

r_c = coefficient of correlation by Concurrent deviation

c = Concurrent deviation number

N = No. of pairs (which is one less than total numbers)

Note : ± signs have been used outside and inside of square root. The reason is that we cannot take square root of minus sign. Therefore, for calculation purposes, if $\left(\frac{2c-N}{N}\right)$ is negative, negative (-) sign inside square root shall be taken so as to have a positive value and if it is positive, the '+' sign inside square root shall be taken so that the value remains positive. Thus, if $\left(\frac{2c-N}{N}\right)$ is negative it will be multiplied by minus sign so that it is converted into a positive sign but again after the square root is taken, it will be multiplied by negative sign and hence the result will be negative. Similarly, if $\left(\frac{2c-N}{N}\right)$ is positive, '+' sign shall be taken inside as well as outside square root. The result is a positive correlation.

Illustration : 13

From the following data relating to Index Number of Supply and Prices. Calculate coefficient of correlation by concurrent deviation method :

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Index No. of Supply	108	120	125	115	115	118	124	124	128	130
Index No. of Prices	100	97	97	99	96	90	92	92	90	88

Solution :

Calculation of Correlation by Concurrent Deviation

Index No. of Supply	Deviation Signs	Index No. of Prices	Deviation Signs	Product of Deviation Signs
108		100		
120	+	97	-	-
125	+	97	=	=
115	-	99	+	-
115	=	96	-	-
118	+	90	-	-
124	+	92	+	+
124	=	92	=	+
128	+	90	-	-
130	+	88	-	-
	N = 9			**C = 2**

$$r_c = \pm\sqrt{\pm\left(\frac{2c-N}{N}\right)} = \sqrt{\pm\left(\frac{2\times 2-9}{9}\right)} = -\sqrt{-\left(\frac{-5}{9}\right)} = -\sqrt{+0.5555} = -0.7453$$

Hence, the coefficient of correlation between Index no. of supply and price is -0.7453.

Merits of Concurrent Deviation Method

(i) It is the simplest mathematical method.

(ii) This method may be used to form a quick idea about the degree of relationship before making use of more complicated methods.

Limitations

(i) Equal weight is given for the changes in same direction, In other words, no difference is made between big or small changes. For example if the change is from 100 to 150 the sign will be plus. If it is from 100 to 500 the sign will also be plus.

(ii) This method is only a rough indicator of presence or absence of correlation.

11.5 Lag and Lead

Correlation is significant when there is a cause and effect relationship among the variables. While calculating coefficient of correlation we assume that out of the two variables one is independent and the other is dependent and there is no significant gap between the cause and its effect. Very often the effect of the cause is not immediate. The change takes place in dependent series after some time. For example change in the value of money does not make immediate effect on prices of commodity. In the same way, the supply of goods may take some time to adjust according to demand. Thus, we find that there is some time gap between cause and effect. This time gap is known as Lag and Lead. When the effect takes place after some time then this is known as Lag or the effect lagging behind. As the cause takes place first so it is leading, regarded as Lead. Thus, here are two words for the same fact.

When there is a time gap between the two variables, this gap must be determined and adjusted accordingly; otherwise, we will not be able to calculate correct coefficient of correlation. After calculating time gap, the dependent series is so adjusted that this gap is eliminated. This is clear from the following illustration.

Illustration : 14

Find if there is any correlation between money in circulation and general price level supposing that the money in circulation affects the wholesale prices in the next year.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Money Circulation	105	111	108	102	107	115	122	116	120	122	128
General Prices	110	113	110	115	109	102	108	122	126	118	120

Solution :

In this question, there is a lag of one year. As such, general prices will be considered from 1996.

Calculation of Coefficient of Correlation

Years	Money in Circulation x	A = 115 dx	dx^2	General Prices y	A = 118 dy	dy^2	dxdy
1995	105	-10	100	113	-5	25	50
1996	111	-4	16	110	-8	64	32
1997	108	-7	49	115	-3	9	21
1998	102	-13	169	109	-9	81	117
1999	107	-8	64	102	-16	256	128
2000	115	0	0	108	-10	100	0
2001	122	+7	49	122	+4	16	28
2002	116	+1	1	126	+8	64	8
2003	120	+5	25	118	0	0	0
2004	122	+7	49	120	+2	4	14
		- 22	**522**		**- 37**	**619**	**398**

$$r = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sqrt{\sum dx^2 \times N - \left(\sum dx\right)^2}\ \sqrt{\sum dy^2 \times N - \left(\sum dy\right)^2}}$$

$$r = \frac{398 \times 10 - (-22 \times -37)}{\sqrt{522 \times 10 - (-22)^2}\ \sqrt{619 \times 10 - (-37)^2}}$$

$$r = \frac{3980 - 814}{\sqrt{5220 - 484}\ \sqrt{6190 - 1369}} = \frac{3166}{\sqrt{4736 \times 4821}}$$

r = +0.6627

Activity B :

State whether following statements are true or false : -

(i) When the variables of two correlated series show changes in the same direction, there is a positive correlation.

(ii) The measurement of perfect Negative correlation is +1.

(iii) Formula for co-variance is $\frac{\sum xy}{N}$

(iv) If coefficient of correlation is less than P.E., then it proves that there is a correlation between the two series.

(v) The sum of rank differences is always one.

(vi) Under normal conditions demand and prices of a commodity are negatively correlated.

(vii) Blood pressure and income of persons are positively correlated.

(viii) Karl Pearson's coefficient of correlation is an absolute measure.

(ix) Coefficient of concurrent deviations can be negative.

(x) Coefficient of correlation must be in the same unit as the original data.

Coefficient of Determination : For proper interpretation of correlation coefficient, coefficient of determination is calculated. This is the square of correlation coefficient.

Coefficient of determination $= r^2$
Coefficient of non determination $= 1-r^2$

11.6 Summary

If the values of two variables vary in such a way that fluctuations in one are accompanied by the fluctuations in the other, these variables are said to be correlated. Correlation is the measure of nature, tendency and limit of connection between two related data series. This analysis contribute to the understanding and locating the critically important variable on which others depend.

The most important ways of classifying correlation are : -

1. Positive and Negative Correlation
2. Linear and Non linear Correlation
3. Simple, Partial and Multiple Correlation

Degree of correlation is studied with the help of coefficient of correlation. On this basis the results of positive and negative correlation may be perfect, limited degree (high, moderate or low degree) or absence of correlation. Important formula to calculate coefficient of correlation are :-

(A) Karl Pearson's Coefficient of Correlation

(i) Individual Series

Direct Method

$$r = \frac{\text{Co-variance of X and Y}}{\sigma_x \times \sigma_y}$$

$$or, r = \frac{\frac{\sum xy}{N}}{\sigma_x \times \sigma_y} \text{ or } \frac{\sum xy}{N \times \sigma_x \times \sigma_y}$$

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

Short-cut Method

$$r = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sqrt{\sum dx^2 \times N - \left(\sum dx^2\right)}\sqrt{\sum dy^2 \times N - \left(\sum dy^2\right)}}$$

(ii) Discrete & Continuous Series

$$r = \frac{\sum fdxdy \times N - \left(\sum fdx \times \sum fdy\right)}{\sqrt{\sum fdx^2 \times N - \left(\sum fdx^2\right)}\sqrt{\sum fdy^2 \times N - \left(\sum fdy^2\right)}}$$

$$P.E. = .6745\frac{1-r^2}{\sqrt{N}}$$

(B) Rank Correlation

$$r_R = 1 - \frac{6\sum D^2}{N\left(N^2-1\right)}$$

$$r_R = \frac{1 - 6\left[\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \ldots\right]}{N\left(N^2-1\right)}$$

(C) Correlation by Concurrent Deviation Method

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

11.7 Key Words

Correlation : A relation between two series or groups of data.

Rank Correlation : The study of correlation of the characteristics expressed by ranks (Qualitative data) is called rank correlation.

Positive Correlation : When both the variable vary in the same direction.

Negative Correlation : When the changes in variables are in opposite direction.

Probable Error : The value which is added to and subtracted from coefficient of correlation to give those limits within which the conclusions of most of the sample should lie.

11.8 Self Assessment Questions

1. What is meant by correlation?
2. Explain the principle of cause and effect. Mention the conditions in which the correlation becomes spurious.
3. Write assumptions of Karl Pearson's coefficient of correlation.
4. Why is Probable Error important? Give formula also.
5. What is Rank Correlation? How will you determine it?
6. What is Correlation of concurrent deviations? How will you measure it?
7. Plot the following data as a scatter diagram and comment over the result:
   Speed of Machine (rpm)

X	10	11	15	13	10	12	13	9	16	15	10
Number of Defectives											
Y	7	8	9	10	7	11	9	7	13	12	7

8. Observe the correlation through graphic method :

Years	1996	1997	1998	1999	2000	2001	2002	2003
Production of Sugar (lakh tons)	31	30	32	35	32	30	34	38
No. of Labourers (in thousand)	80	77	82	85	84	82	84	86

9. (a) From the following data compute coefficient of correlation between X and Y :

	X-Series	Y-Series
No. of Items	15	15
Arithmetic Mean	25	18
Squares of Deviations from Mean	130	135

Summation of product of deviations of X and Y series from their respective means = 115.

(r = .86)

(b) Find out the number of items if :

$r = .4$, $\Sigma xy = 80$, $\sigma_y = 5$, $\Sigma x^2 = 80$ (N = 8)

10. Calculate Karl Pearson's coefficient of correlation between the ages of husbands and age of wives:

Husband's Age	23	27	28	29	30	31	33	35	36	39
Wife's Age	18	22	23	24	25	26	28	29	30	32

(r = 0.9955)

11. From the following table giving the distribution of students and also regular players among them according to age-groups, find the correlation between age and playing habit.

Age	No. of Students	Regular Players
14-15	200	150
15-16	270	162
16-17	400	200
17-18	360	180
18-19	400	180
19-20	200	80

(r = -.94)

12. From the following data, find out if there is any relationship between density of population and death rate :

Zones	Area (Sq. Km.)	Population	No. of Deaths
P	120	24	288
Q	150	75	1125
R	80	48	768
S	50	40	720
T	200	50	650

(r = 0.997)

13. Do you find any correlation between age and intelligence in the following data?

Marks	Age in years				
	17	18	19	20	21
25-30	3	2	---	---	---
20-25	---	5	4	---	---
15-20	---	---	7	10	---
10-15	---	---	---	3	2
5-10	---	---	---	3	1

(Ans. : r = -.837)

14. Ten competitors in a beauty competition are ranked by three judges in the following order. Use the rank correlation coefficient to discuss which pair of judges has the nearest approach to common tastes in beauty.

First Judge	2	3	6	9	4	10	7	5	8	1
Second Judge	10	7	5	1	2	4	6	8	9	3
Third Judge	3	2	10	8	1	9	5	7	6	4

(r_R = I & III Judge

15. Calculate correlation coefficient by the concurrent deviation method from the following data :

X	:	70	72	75	75	78	80	82	83	84	84	90
Y	:	48	49	52	52	50	53	54	55	60	61	62

(Ans. : r_c = +0.77)

11.7 Reference Books

1. Gupta, S.P., Statistical Methods.
2. Sharma, J.K., Business Statistics.
3. Agarwal, D.R., Business Statistics.
4. Pinnai & Bhagwati, Statistical Methods.
5. Nagar, K.N., Statistical Methods.
6. Yadav, Jain, Mittal, Statistical Methods.

Unit - 12 Regression Analysis

Structure of Unit:

12.0 Objectives

After completing this unit, you will be able to : -

- Define regression analysis.
- Evaluate the difference between correlation and regression analysis.
- Explain the meaning of regression lines and their functions.
- Explain various methods of constructing regression lines & equations.
- Determine regression coefficients and with the help of these regression coefficient, will be able to calculate correlation coefficient.
- Assess the importance of standard error of estimate.

12.1 Introduction

Correlation only indicates the degree and direction of relationship between two variables. It does not, necessarily connote a cause-effect relationship. Even when there are grounds to believe the casual relationship exists, correlation does not tell us which variable is the cause and which, the effect. For example, rainfall and production of rice will generally be found to be correlated, but if we are interested to know the estimate rainfall for certain amount of production or vice versa, it will not be answered by correlation. Regression analysis clearly indicates the cause and effect relationship which enable us to make estimates of one variable from another variable. A better understanding of this casual relationship helps in prediction and influencing and control over the future course of a given phenomenon. The variable constituting cause is taken as independent variable and the variable constituting the effect is taken as dependent variable.

12.2 Meaning and Definition of Regression

In 1877, Sir Francis Galton originally introduced the term regression, in his Research Paper "Regression towards Mediocrity in Hereditary Stature". He had made a study of the height of one thousand fathers and sons. He drew an interesting conclusion that tall fathers had tall sons and short statured fathers had short statured sons. In other words, there was a high degree positive correlation between them. But, the height of fathers had more deviations from the general average, while the height of sons had less deviations from the general average. Thus there was a tendency to move towards the mediocrity. Galton described the average relationship between the height of fathers and their sons as the Line of Regression.

Regression tell us about the average relationship between two variables from which estimations can be done.

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

According to **Walls and Roberts**, "It is often more important to find out what the relationship actually is, in order to estimate or predict one variable (the dependent variable); and the statistical technique appropriate to such a case is called regression analysis."

Taro Yamana says, "One of the frequently used techniques in economics and business research to find a relation between two or more variables that are related casually, is regression analysis."

Thus, regression analysis is a mathematical measure of the average relationship between a series of two or more variable and generally used to predict the value of one variable on the basis of another variable.

12.3 Utility of Regression Analysis

Regression Analysis is highly useful in almost all sciences - natural or social. The following are some of the important functions or uses of regression analysis : -

1. Regression analysis can be used in all those areas in which there is a tendency to move towards general mean in variables of two or more related series.

2. Forecasting can be done of one dependent variable in relation with the other independent variable by regression analysis.

3. Degree and direction of correlation can be estimated by regression analysis. It is calculated with the help of regression whether there exists perfect correlation or absence of correlation between two variables.

4. Regression analysis can be used as a control tool in economic and business field. Not only decision making becomes easy on the basis of this technique but decisions can also be tested on practical basis.

Difference between Correlation and Regression

Pointing out the difference between Correlation and Regression, W.Z. Hursh says, "Correlation analysis examines the intensity of co-variation in two or more incidents where as regression analysis provides capacity to estimate by measuring the nature and degree of this relation."

Difference between Correlation and Regression

Base	Correlation	Regression
1. Degree and nature of relationship	Correlation tells the degree of co-variance between two or more variables.	Using relationship between known variable and unknown variable to estimate or predict the unknown variable is regression analysis.
2. Cause and effect	Correlation is unable to tell which series is the cause and which is the effect, despite high degree relation between two series.	In regression the given independent variable is the 'cause' and dependent variable is the 'effect'.
3. Effect of change in scale	Coefficient of correlation is unaffected by the change of origin and scale.	Regression coefficient is affected by change in scale.

Types of Regression Analysis

In regression analysis we study mean variation in values of a variable due to certain variation in the values

of the other variable. The first variable whose value is known is called independent variable while the other, whose value is unknown is called dependent variable. These variables are denoted by X and Y. Regression analysis is of two types : -

1. Linear and Curvilinear regression;
2. Simple and Multiple regression.

1. Linear and Curvilinear regression : - When variables of two related series X - Y are plotted on graph, this gives us a scatter diagram. If two best fit lines are drawn between the middle points of this scatter diagram, these very lines are called regression lines. If these lines are straight, it is called linear regression. If these lines are simple and in the form of a smooth curve, it is called curvilinear regression.

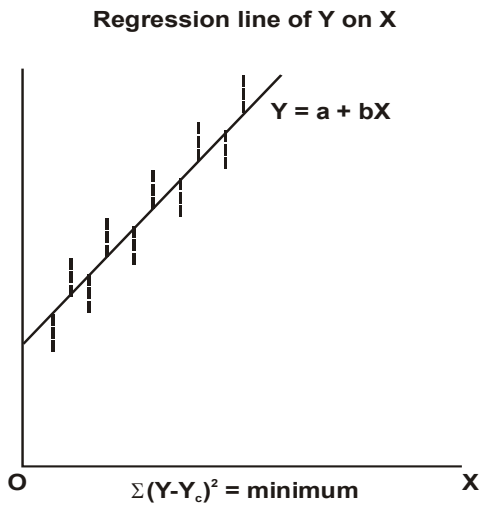
2. Simple and Multiple regression : - When we study average relationship of only two related variables, it is called simple regression. The known variable is independent, the variable whose value is to be estimated is called dependent variable. When regression analysis method is used to examine relationship of more than two variables, it is called multiple regression. In this chapter we shall discuss simple linear regression only.

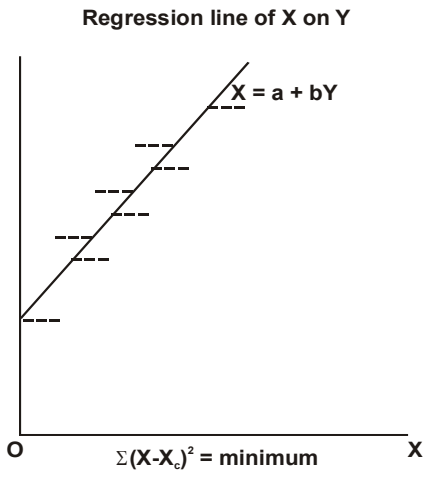
12.4 Regression Lines

The lines of best fit drawn to show the mutual relationship between X and Y variables are known as Regression Lines. For two mutually related series, we have two regression lines, one representing regression of X on Y and other of Y on X. The line representing regression of X on Y assumes Y as an independent variable and X as a dependent variable. This line gives most probable value of X for the given value of Y. In the same way, the second line represents the regression of Y on X. This is drawn by assuming X as an independent variable and Y as a dependent variable. This line gives the most probable value of Y for the given value of X.

The regression lines are drawn on least square assumption. Under this method the line is so drawn that the total of squared deviations from different points to the line is minimum. The deviations from the point to the line can be measured by two ways, first horizontally that is parallel to X axis and second, vertically that is parallel to Y axis. To make the sum of squared deviations minimum from both sides, two lines are essential. Hence, we need two regression lines for two series.

When there is perfect correlation (+1 or -1) between two series, the two regression lines will overlap and there will be only one regression line.





12.4.1 Functions of Regression Lines

Following are the two most important functions of regression lines : -

1. Best Estimate : We can have the best estimates from regression lines. Variable X can be estimated from regression line of X on Y, and variable Y can be estimated from regression line of Y on X.

2. Extent and Direction of Correlation : The extent and direction of correlation can be known by regression lines in the following ways :

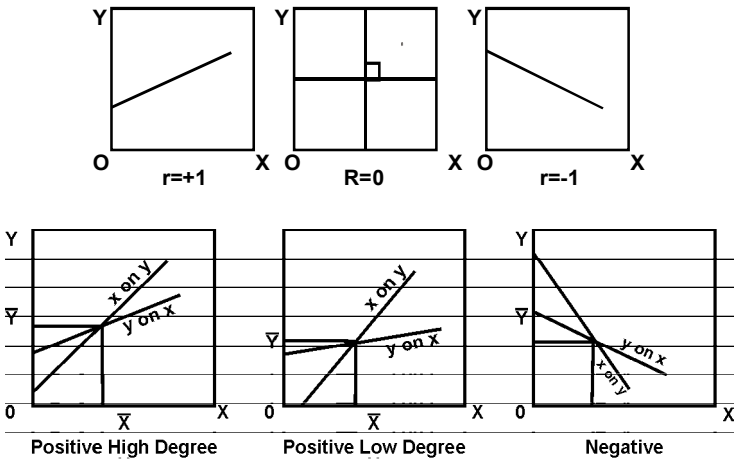
(i) Positive Correlation : When both the regression lines show an upward trend from left hand corner to right hand upward side then positive correlation is found between X and Y series.

(ii) Negative Correlation : When both the regression lines show a downward trend from upper left hand side to downward right hand side then negative correlation is found between X and Y series.

(iii) Absence of Correlation : If these lines intersect each other at right angle (i.e. 90^0), there will be absence of correlation between two series. In other words, degree of correlation will be zero.

(iv) Perfect Correlation : If all the marked points on the graph are in a straight line and the two lines overlap each other, there will be perfect correlation in the series.

(v) Limited Correlation : Regression lines cut each other at average points. If we draw perpendiculars to Y axis and X axis from the intersection point of regression lines, then we will find average points of X and Y series. The nearer the regression lines to each other the higher the degree of correlation and the farther the regression line from each other, the lesser is the degree of correlation. These points are explained through the following graphs:



12.4.2 Methods of Constructing Regression Lines

Regression lines can be drawn by two methods :

(i) **Free hand method : -** Regression lines drawn by two persons by free hand method will not be the same. Different persons will draw different lines. Hence, regression lines are drawn on the basis of regression equations.

(ii) **Regression equations method : -** Regression Equations are algebraic form of regression lines. Like lines, equations are also of two types : -

a) **Regression equation of X on Y : -** This equation is used to study the changes in values of the dependent variable X for given values of independent variable Y. Its original form is as follows:

$$X = a + bY$$

b) Regression equation of Y on X : - This equation is used to study the changes in values of the dependent variable Y for given values of independent variable X. It is written in the form of : -

$$Y = a + bX$$

Analysis of equation : - In the above equation X and Y are variables, and a and b are constants. Values of these constants are fixed. These constants are called parameters of equation. Constant 'a' determines

the values of dependent variable when values of independent variable is zero and thus determines the height or level of regression line. If 'a' is positive, the line will begin above the point of origin at Y axis. If parameter 'a' is negative, the line will begin below the point of origin and if 'a' is zero, the line will begin from the point of origin. Parameter 'b' tells about the slope of regression line. It also tells the change in dependent variable which happens due to the change in single unit of independent variable. If 'b' has positive value, the line will go upward from the left to the right; and if it is negative, the slope will be downward from upper left hand side to downward right hand side. The value of 'b' like 'a' cannot be zero. 'b' is also called regression coefficient of regression equations.

There are two ways to know parameters 'a' and 'b' : -

(A) Mean based;
(B) Least squares method.

(A) Mean Based : - Equation based on mean are as follows : -

(a) When the values of mean, standard deviation and correlation coefficient are already known:

(i) Regression Equation of X on Y

$X = a + bY$

$X = \left(\overline{X} - b\overline{Y}\right) + bY \quad \left[\therefore\ a = \overline{X} - b\overline{Y}\ \right]$

$or\ \left(X - \overline{X}\right) = bY - b\overline{Y}$

$or\ \left(X - \overline{X}\right) = b_1\left(Y - \overline{Y}\right) \qquad \therefore \left[b_1 = r\frac{\sigma_x}{\sigma_y}\right]$

$or\ \left(X - \overline{X}\right) = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$

(ii) Regression Equation of Y on X

$Y = a + bX$

$or\ \ Y = \left(\overline{Y} - b\overline{X}\right) + bX \qquad \left[\therefore\ a = \overline{Y} - b\overline{X}\right]$

$or\ \ Y - \overline{Y} = bX - b\overline{X}$

$or\ Y - \overline{Y} = b_2\left(X - \overline{X}\right) \qquad \therefore \left[b_2\ r = \frac{\sigma_y}{\sigma_x}\right]$

$\left(Y - \overline{Y}\right) = r\ \frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$

Here,

(i) b and b are regression coefficients of X on Y and Y on X respectively.
ii) $\frac{1}{X}$ and $\frac{2}{Y}$ are mean values of X and Y series respectively.
iii) σ_x and σ_y are values of standard deviations of X and Y series respectively.
iv) r is the correlation coefficient of X and Y.

The estimated value of X can be calculated from regression equation of X on Y; in the same way the estimate of Y can be found out by regression equation Y on X, which will be clear from the following example :

Illustration : 1

The following information about advertisement and sales are available :

	Advertisement exp. (X) (Rs. crores)	Sales (Y) (Rs. crores)
Mean	20	120
S.D.	5	25
Correlation coefficient =	0.8	

Calculate the two regression equations and estimate the cost of advertisement expenses when sales is Rs. 200 crores.

Solution :

Given : $\overline{X} = 20;\ \overline{Y} = 120;\ \sigma_x = 5;\ \sigma_y = 25\ and\ r = 0.8$

(i) Regression Equation of X on Y :

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$$

$\therefore$ X -20=.8$\frac{5}{25}$$(Y-120)$

or X -20=.16 $(Y-120)$

or X -20 =.16Y - 19.20

or X =.16Y+20 - 19.20

or X =.16Y+.8

(ii) Regression Equation of Y on X :

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

$\therefore$ Y - 120 =.8$\frac{25}{5}$$(X-20)$

or Y - 120 = 4 $(X-20)$

or Y - 120 = 4X - 80

or Y = 4X + 120 - 80

or Y = 4X + 40

Thus regression equations are :

X = .16Y + 0.8

and Y = 4X + 40

now if Y = 200 then X =

X = .16y +0.8

X = .16 x 200 + 0.8

X = 32 + 0.8

X = 32.8

(b) When deviations are taken from actual mean : -

(i) Regression Equation X on Y

$$\left(X - \overline{X}\right) = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$$

$$\left(X - \overline{X}\right) = \frac{\sum xy}{N\sigma_x \times \sigma_y} \times \frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$$

$$\left(X - \overline{X}\right) = \frac{\sum xy}{N\sigma_y^2}\left(Y - \overline{Y}\right)$$

$$\left(X - \overline{X}\right) = \frac{\sum xy}{\sum y^2}\left(Y - \overline{Y}\right)$$

(ii) Regression Equation Y on X

$$\left(Y - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$$

$$\left(Y - \overline{Y}\right) = \frac{\sum xy}{N\sigma_x \times \sigma_y} \times \frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

$$\left(Y - \bar{Y}\right) = \frac{\sum xy}{N\sigma_x^2} \left(X - \bar{X}\right)$$

$$\left(Y - \bar{Y}\right) = \frac{\sum xy}{\sum x^2} \left(X - \bar{X}\right)$$

Here,

Σxy = Sum of the product of deviations of X and Y from their respective actual means.

Σx^2 and Σy^2 = Totals of squares of such deviations from actual means of X and Y series respectively.

Illustration : 2

From the following data calculate Regression Equations by taking deviations from means of X and Y Series :

X	3	4	6	8	9
Y	5	8	7	6	9

Solution :

Calculation of Regression Equations

X	$(X - \bar{X})$ **x**	$\mathbf{x^2}$	**Y**	$(Y - \bar{Y})$ **y**	$\mathbf{y^2}$	**xy**
3	-3	9	5	-2	4	6
4	-2	4	8	+1	1	-2
6	0	0	7	0	0	0
8	+2	4	6	-1	1	-2
9	+3	9	9	+2	4	6
30	**0**	**26**	**35**	**0**	**10**	**8**

$$\bar{X} = \frac{\sum x}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{35}{5} = 7$$

(i) Regression Equations of X on Y

$$\left(X - \bar{X}\right) = \frac{\sum xy}{\sum y^2} \left(Y - \bar{Y}\right)$$

$$(X - 6) = \frac{8}{10} (Y - 7)$$

$(X - 6) = .8 (Y - 7)$

$X = .8y - 5.6 + 6$

$X = .4 + .8y$

(ii) Regression Equations of Y on X

$$\left(Y - \bar{Y}\right) = \frac{\sum xy}{\sum x^2} \left(X - \bar{X}\right)$$

$$(Y - 7) = \frac{8}{26} (X - 6)$$

$Y - 7 = .31 (X - 5)$

$Y = 7 + .31x - 1.55$

$Y = 5.45 + .31x$

Activity : A

With the help of the following data give the two regression equations, plot them on graph paper and show the degree of correlation :

$\bar{X} = 30;\ \bar{Y} = 25;\ \sigma_X = 4.05;\ \sigma_Y = 3.7;\ r = .8$

(c) Determination of Regression Equation when deviations are taken from assumed mean : -

(i) Regression Equation of X on Y

$$X - \overline{X} = b_{xy}\ (Y - \overline{Y})$$

$$\left(X - \overline{X}\right) = \frac{\sum dxdy - \frac{\left(\sum dx \times \sum dy\right)}{N}}{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}}\left(Y - \overline{Y}\right)$$

(ii) Regression Equation of Y on X

$$Y - \overline{Y} = b_{yx}\ (X - \overline{X})$$

$$\left(Y - \overline{Y}\right) = \frac{\sum dxdy - \left[\frac{\left(\sum dx.\sum dy\right)}{N}\right]}{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}}\left(X - \overline{X}\right)$$

Here,

Σdxdy = Summation of the products of corresponding deviations of X and Y from their respective assumed means.

Σdx and Σdy = Sum of deviations of X and Y values from their assumed means.

Σdx^2 and Σdy^2 = Sum of squares of deviations of X and Y from their assumed means.

Illustration : 3

From the following data obtain two regression equations : -

X	2	4	6	8	10
Y	5	7	8	9	11

Solution :

Calculation of Regression Equations

X	dx (A=6)	dx^2	Y	dy (A=8)	dy^2	dxdy
2	-4	16	5	-3	9	12
4	-2	4	7	-1	1	2
6	0	0	8	0	0	0
8	2	4	9	1	1	2
10	4	16	11	3	9	12
Total	**0**	**40**	**40**	**0**	**20**	**28**

$$\overline{X} = A + \frac{\sum dx}{N}$$

$$\overline{X} = 6 + \frac{0}{5} = 6$$

Regression Coefficient of X on Y

$$b_{xy} = \frac{\sum dxdy - \frac{\sum dx.dy}{N}}{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}}$$

$$\overline{Y} = A + \frac{\sum dy}{N}$$

$$\overline{Y} = 8 + \frac{0}{5} = 8$$

Regression Coefficient of Y on X

$$b_{yx} = \frac{\sum dxdy - \frac{\sum dx.dy}{N}}{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}}$$

$$b_{xy} = \frac{28 - \frac{0.0}{5}}{20 - \frac{(0)^2}{5}}$$

$$b_{xy} = \frac{28}{20} = 1.4$$

$$X - \overline{X} = bxy\left(Y - \overline{Y}\right)$$

$$X - 6 = 1.4(Y - 8)$$

$$X - 6 = 1.4Y - 11.2$$

$$X = 1.4Y - 11.2 + 6$$

$$X = 1.4Y - 5.2$$

$$b_{yx} = \frac{28 - \frac{0.0}{5}}{40 - \frac{(0)^2}{5}}$$

$$b_{yx} = \frac{28}{40} = .7$$

$$Y - \overline{Y} = byx\left(X - \overline{X}\right)$$

$$Y - 8 = .7(X - 6)$$

$$X - 8 = .7X - 4.2$$

$$Y = .7X - 4.2 + 8$$

$$Y = .7X + 3.8$$

(iv) If the regression equations are calculated from values of X and Y, where step deviations are taken : -

(i) Regression Equation of X on Y

$$\left(X - \overline{X}\right) = \frac{\sum dxdy - \left(\frac{\sum dx.\sum dy}{N}\right)}{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}} \times \frac{i_x}{i_y}\left(Y - \overline{Y}\right)$$

(ii) Regression Equation of Y on X

$$\left(Y - \overline{Y}\right) = \frac{\sum dxdy - \left(\frac{\sum dx.\sum dy}{N}\right)}{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}} \times \frac{i_y}{i_x}\left(X - \overline{X}\right)$$

Here,

i_x = Step deviation of X variable

i_y = Step deviation of Y variable

(v) Regression Equations in Bi-variate Grouped Frequency Distribution

(a) Regression Equations of X on Y

$$\left(X - \overline{X}\right) = b_{xy}\left(Y - \overline{Y}\right)$$

$$\left(X - \overline{X}\right) = \frac{\sum fdxdy - \left(\frac{\sum fdx.\sum fdy}{N}\right)}{\sum fdy^2 - \frac{\left(\sum fdy\right)^2}{N}} \times \frac{i_x}{i_y}\left(Y - \overline{Y}\right)$$

(b) Regression Equations of Y on X

$$\left(Y - \overline{Y}\right) = b_{yx}\left(X - \overline{X}\right)$$

$$\left(Y-\overline{Y}\right)=\frac{\sum fdxdy-\left(\frac{\sum fdx.\sum fdy}{N}\right)}{\sum fdx^2-\frac{\left(\sum fdx\right)^2}{N}}\times\frac{i_y}{i_x}\left(X-\overline{X}\right)$$

Illustration : 4

The following table gives the number of students having different heights and weights : -

Heights (in inches)	Weights (in lbs)				
	80-90	90-100	100-110	110-120	Total
50-55	6	10	4	--	20
55-60	4	10	10	1	25
60-65	4	8	15	8	35
65-70	4	3	2	11	20
Total	**18**	**31**	**31**	**20**	**100**

On the basis of the above data, calculate regression equations.

Solution :

Calculation of Regression Equations

Height (X)		Weight (Y)					f		fdx	fd^2x
			80-90	90-100	100-110	110-120				
			85	95	105	115				
		dy \ dx	-1	0	1	2				
50-55	52.5	-2	2 6 12	0 10 0	-2 4 -8	-4 - 0	20		-40	80
55-60	57.5	-1	1 4 4	0 10 0	-1 10 -10	-2 1 -2	25		-25	25
60-65	62.5	0	0 4 0	0 8 0	0 15 0	0 8 0	35		0	0
65-70	67.5	1	-1 4 -4	0 3 0	1 2 2	2 11 22	20		20	20
		f	18	31	31	20	100 N		-45 Σfdx	125 fd^2x
		fdy	-18	0	31	40	53 Σfdy			
		fdy^2	18	0	31	80	129 Σfd^2y			
		fdxdy	12	0	-16	20	16 Σfdxdy			

Mean :

(i) $\bar{X} = A_x + \frac{\sum fdx}{N} \times i = 62.5 + \frac{-45}{100} \times 5$

$= 62.5 - 2.25 = 60.25$

(ii) $\bar{Y} = A_y + \frac{\sum fdy}{N} \times i = 95 + \frac{53 \times 10}{100}$

$= 95 + 5.3 = 100.3$

Regression Coefficients

(i) $b_{xy} = \frac{\sum fdxdy - \frac{\sum fdx \times \sum fdy}{N}}{\sum fdy^2 - \frac{(\sum fdy)^2}{N}} \times \frac{i_x}{i_y}$

$= \frac{16 - \left(\frac{-45 \times 53}{100}\right)}{129 - \frac{(53)^2}{100}} \times \frac{5}{10}$

$= \frac{16 + 23.85}{129 - 28.09} \times \frac{5}{10} = \frac{199.25}{1009.1}$

$b_{xy} = .197$

(ii) $b_{yx} = \frac{\sum fdxdy - \frac{\sum fdx - \sum fdy}{N}}{\sum fdx^2 - \frac{(\sum fdx)^2}{N}} \times \frac{i_y}{i_x}$

$= \frac{16 - \left(\frac{-45 \times 53}{100}\right)}{125 - \frac{(-45)^2}{100}} \times \frac{10}{5}$

$= \frac{16 + 23.85}{125 - 20.25} \times \frac{10}{5}$

$= \frac{398.5}{523.75}$

$b_{yx} = .761$

Regression Equation of X on Y

$X - \bar{X} = b_{xy} \left(Y - \bar{Y}\right)$

$= X - 60.25 = .197 \ (Y - 100.3)$

$= X - 60.25 = .197Y - 19.76$

$X = 60.25 - 19.76 + .197Y$

$X = 40.49 + .197Y$

Regression Equation of Y on X

$Y - \bar{Y} = b_{yx} \left(X - \bar{X}\right)$

$Y - 100.3 = 0.761 \ (X - 60.25)$

$Y - 100.3 = 0.761 \ X - 45.85$

$Y = 100.3 - 42.85 + 0.761 \ X$

$Y = 54.45 + 0.761X$

B. Regression Equations by Least Squares Method

As we have already explained that the regression lines are those best fit lines which are drawn on least square assumptions. For drawing these lines mathematical equations are used. These lines are based on these two important characteristics of Least Square : -

(i) $\Sigma(Y-Y_c) = 0$: The sum of deviations of given values and estimated values of dependent variable are always equal to zero.

(ii) $\Sigma(Y-Y_c)^2$ = minimum : The sum of the squares of deviations from regression line is less than the sum of square of deviations from any other line.

For calculating regression equations, under least squares method we have to calculate the parameters 'a' and 'b' by two normal equations. The values of 'a' and 'b' are calculated in the following manner : -

(i) Regression Equations of X on Y

$X = a + bY$

Normal equations

$\Sigma X = Na + b\Sigma Y$ ..... (i)
$\Sigma XY = a\Sigma Y + b\Sigma Y^2$ ..... (ii)

(ii) Regression Equation of Y on X

$Y = a + bX$

Normal equations

$\Sigma Y = Na + b\Sigma X$ ..... (i)
$\Sigma XY = a\Sigma X + b\Sigma X^2$ ..... (ii)

On the basis of the above equations we will find the values of 'a' and 'b' and from them regression equations may be formed by the help of basic regression equations.

Here, ΣX, ΣY, ΣX^2, ΣY^2 and ΣXY are the sums of different values, their squares and product respectively.

Illustration : 5

From the following data, calculate regression equations by least squares method :

X	1	3	5	6	5
Y	2	4	6	8	10

Solution :

Calculation of Regression Equations

X	Y	X^2	Y^2	XY
1	2	1	4	2
3	4	9	16	12
5	6	25	36	30
6	8	36	64	48
5	10	25	100	50
20	**30**	**96**	**220**	**142**

Regression Equation of X on Y

$\Sigma X = Na + b\Sigma Y$ ..... (i)

$\Sigma XY = a\Sigma Y + b\Sigma Y^2$ ..... (ii)

by putting the values -

$20 = 5a + 30\ b$ ..... (i)

$142 = 30\ a + 220\ b$ ..... (ii)

Multiplying equation (i) by 6 and subtracting equation (ii) from it -

$120 = 30\ a + 180\ b$

$142 = 30\ a + 220\ b$

(-) (-) (-)

$-\ 22 = -40\ b$

$$\text{or } b = \frac{22}{40} = .55$$

Putting the value of b in equation (i)

$20 = 5a + 30 \times .55$
$20 = 5a + 16.5$
$20 - 16.5 = 5a$
$5a = 3.5$

$$a = \frac{3.5}{5}$$

$a = 0.7$

Hence, regression equation of X on Y is

$X = a + bY$
$X = 0.7 + .55Y$

Regression equation of Y on X

$\Sigma Y = Na + b\Sigma X$ ..... (i)
$\Sigma XY = a\Sigma X + b\Sigma X^2$ ..... (ii)

by putting the values

$30 = 5\ a + 20\ b$ ..... (i)
$142 = 20\ a + 96\ b$ ..... (ii)

Multiplying equation (i) by 4 and subtracting equation (ii) from it.

$120 = 20\ a + 80\ b$
$142 = 20\ a + 96\ b$

(-) (-) (-)

$-\ 22 = -\ 16\ b$

$$\text{or } b = \frac{22}{16} = 1.375$$

Putting the value of b in equation (i) : -

$30 = 5a + 20 \times 1.375$
$30 = 5a + 27.5$
$30 - 27.5 = 5a$
$2.5 = 5a$

$$a = \frac{2.5}{5}$$

$a = 0.5$

Hence, regression equation of Y on X is

$Y = a + bX$
$Y = 0.5 + 1.375X$

12.5 Regression Coefficients

1. Regression Coefficient of X on Y : Regression coefficient of X on Y is the measure of the slope of the regression line of X on Y. It tells us that with change of one unit in Y, what change will be in variable X. The notation given to regression coefficient of X on Y is b_{xy} or b_1. It can be calculated by the following formula :

(i) $$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

(ii) $$b_{xy} = \frac{\sum xy}{\sum y^2}$$

(iii) $$b_{xy} = \frac{\sum dxdy - \left(\frac{\sum dx.\sum dy}{N}\right)}{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}}$$

(iv) $$b_{xy} = \frac{\sum dxdy - \left(\frac{\sum dx.\sum dy}{N}\right)}{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}} \times \frac{i_x}{i_y}$$

2. Regression Coefficient of Y on X : This coefficient is the algebric measurement of the slope of regression line Y on X. It tells us that with change of one unit in X, what change will be in variable Y. This coefficient is denoted by b_{yx} or b_2 and can be formulated as follows : -

(i) $$b_{yx} = r\frac{\sigma_y}{\sigma_x};$$

(ii) $$b_{yx} = \frac{\sum xy}{\sum x^2};$$

(iii) $$b_{yx} = \frac{\sum dxdy - \left(\frac{\sum dx.\sum dy}{N}\right)}{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}}$$

(iv) $$b_{yx} = \frac{\sum dxdy - \left(\frac{\sum dx.\sum dy}{N}\right)}{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}} \times \frac{i_y}{i_x}$$

Determination of Coefficient of Correlation from Regression Coefficients

We can determine the coefficient of correlation with the help of both regression coefficients. Actually, coefficient of correlation is the geometric mean of the two regression coefficients. In other words, coefficient of correlation is the square root of products of regression coefficients. It is clear from the following formula:

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$\because b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad and \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$hence, \quad r = \sqrt{r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x}} \quad or = \sqrt{r^2} \quad = \; r$$

$$so, \quad r = \sqrt{b_{xy} \; . \; b_{yx}}$$

Interpretation of regression coefficients : -

(i) Product of both regression coefficients cannot be more than 1.

(ii) If one regression coefficient is more than 1, the value of the other regression coefficient will be so much less than 1 as by multiplying both, the product will not exceed 1. For example, if $b_{xy} = 1.5$ and $b_{yx} = .75$, then r = 1.06 which is not possible, because r can never be more than 1.

(iii) Both regression coefficient either would be positive or negative. In any condition they cannot bear different signs.

(iv) If both the coefficients are positive correlation coefficient will be positive, if both the coefficients are negative sign, the coefficient of correlation will also be negative.

(v) In any regression coefficient four values $b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad or \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$ are used, hence if we know three values, than we can compute the fourth.

Illustration : 6

a) For a distribution the following measures are given :

Mean $\overline{X} = 26.6$, $\overline{Y} = 13.5$

Regression coefficient X on Y = 0.2, Y on X = 1.5

Find out :

(i) Most likely value of Y when X = 40
(ii) Coefficient of correlation.

b) From the following data

$X = 0.64Y$, $Y = 0.81X$, $\sigma_x = 4$

Calculate :

(i) Coefficient of correlation
(ii) Standard deviation of Y series

Solution :

a) (i) Regression Equation of Y on X

$Y = a + bX$

$$\left(Y - \overline{Y}\right) = b_2\left(X - \overline{X}\right)$$

$(Y-13.5)=1.5(X-26.6)$

$(Y-13.5)=1.5X-39.9$

$Y=13.5-39.9+1.5X=-26.4+1.5X$

By substituting the value of X

$Y=-26.4+1.5\times 40$

$\therefore Y=33.6$

(ii) $r=\sqrt{b_1\times b_2}=\sqrt{.2\times 1.5}=\sqrt{0.3}=.548$

b) **(i)** The given values represents : -

$b_{xy}=0.64$, $b_{yx}=0.81$

So, $r=\sqrt{b_{xy}\times b_{yx}}=\sqrt{0.64\times 0.81}=0.72$

(ii) $b_{xy}=r\frac{\sigma_x}{\sigma_y}$, by putting known values

$0.64=0.72\frac{4}{\sigma_y}$

$=0.64\sigma_y=0.72$ x 4

$\sigma_y=\frac{0.72\times 4}{0.64}$

$\sigma_y=4.5$

Activity B

Show the coefficient of correlation is the geometric mean between regression coefficients. If the sign of a regression coefficient is known, how would you find the sign of the coefficient of correlation? If one of the regression coefficients is negative, what type of variation would you expect in the original series of pairs of observations?

_____

_____

_____

12.6 Calculation of Mean Values and Regression Coefficient From Given Equations

When the regression equations of X on Y and Y on X are known, the regression coefficient can easily be separated from them as the values of 'b' are the value of regression coefficients. But sometimes regression equations are given in such a way that it is not possible to know which equation is of X on Y and which equation is of Y on X. In this condition any one of the equations is presumed as X on Y and the other one as Y on X regression equation. After this assumption the regression coefficients are calculated. If the product of b_{xy} and b_{yx} comes to greater than 1 or signs of coefficients comes different (one positive other negative), it means we have assumed equations wrongly. Hence, we should now change our assumption. By changing the assumptions the value of correlation coefficient calculated by regression coefficients will be less than one or at the most one.

For calculating the mean values of X and Y, the values of X and Y should be found out from the given equations. The values of X and Y will be $\overline{X}$ and $\overline{Y}$ respectively.

Illustration : 7

For certain X and Y series which are correlated, the two lines of regression are as given below :

5X - 6Y + 90 = 0 ..... (i)
15X - 8Y - 130 = 0 ..... (ii)

(i) Find the means of two series and the correlation coefficient.
(ii) Find which is regression equation of Y on X and which is X on Y.

Solution :

(i) Calculation of Mean :

5X - 6Y = -90 ..... (i)
15X - 8Y = 130 ..... (ii)

Multiply equation (i) by 3 :

15X - 18Y = -270 ..... (iii)
15X - 8Y = 130 ..... (iv)
- + -

By subtracting

-10Y = -400
or Y = 40

By substituting the value of Y = 40

5X - 6(40) = -90 or 5X = +240 - 90
5X = 150 or X = 30

Thus $\overline{X} = 30$, $\overline{Y} = 40$

(ii) Calculation of Coefficient of Correlation :

It is assumed that equation (i) is Y on X. Therefore b_{yx} is :

5X - 6Y + 90 = 0

-6Y = -5X - 90

$$Y = \frac{-5}{-6}X \quad \therefore b_{yx} = \frac{5}{6} \; or \; .83$$

Now the second equation is assumed as X on Y. Therefore b_{xy} is :

15X = +8Y + 130 or 15X = 8Y

$$\therefore X = \frac{8}{15}Y$$

$$or \; b_{xy} = \frac{8}{15} = .53$$

$$r = \sqrt{b_{xy} \times b_{yx}} \text{ or } \sqrt{.83 \times .53} = 0.663$$

Thus the coefficient of correlation is less than one, therefore the above assumption is correct.

Illustration : 8

Following are the two regression equations : X + 3Y - 09 = 0 and 3X + 2Y - 10 = 0 and $\sigma^2_x = 16,$ then find out $\overline{X}$, $\overline{Y}$, $\sigma_y{}^2 and\ r$

Solution :

(i) Calculation of Mean Values

X + 3Y = 9 ..... (i)
3X + 2Y = 10 ..... (ii)

Multiplying equations (i) by 3 and subtracting equation (ii) from it.

3X + 9Y = 27

3X + 2Y = 10
- - -

7Y = 17

or $Y = \frac{17}{7} = 2.43$

By substituting 2.43 in equation (i)

$X + 3 \times 2.43 = 9\ or\ X = 9 - 7.29;$

$So,\ \overline{X} = 1.71,\ \overline{Y} = 2.43$

(ii) For calculation of regression coefficient it is not known as to which equation is of X on Y, therefore, by assuming (i) equation X+3Y-9=0 as X on Y

X = -3Y+9, hence b_{xy} = -3

Now, assuming (ii) equation 3X+2Y-10=0 as Y on X

$2Y = 10 - 3X\ or\ Y = \frac{10}{2} - \frac{3}{2}X\ \therefore\ b_{xy} = -1.5$

$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{-3 \times -1.5} = \sqrt{4.5} = -1$

The coefficient of correlation cannot be greater than 1 and here it is more than 1. It means our assumptions about equations are wrong. Therefore, by assuming (ii) equation as X on Y and (i) equation as Y on X.

(i) Regression Equation X on Y

3X = 10 - 2Y

$X = \frac{10}{3} - \frac{2}{3}Y$

$\therefore\ b_{xy} = -.67$

(ii) Regression Equation Y on X

3Y = 09-X

$Y = \frac{09}{3} - \frac{1}{3}X, \quad \therefore\ b_{yx} = -\frac{1}{3}$

$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{-.67 \times -\frac{1}{3}} = -\sqrt{.223} = -.47$

For calculating σy^2 we will use the formula of b_{xy}

$$b_{xy} = \frac{r \times \sigma_x}{\sigma_y}$$

$$-.67 = \frac{-.47 \times 4}{\sigma_y}$$

$$-.67\sigma_y = -.47 \times 4$$

$$\sigma_y = \frac{-.47 \times 4}{-.67} = 2.8$$

$$\sigma_y^2 = (2.8)^2 = 7.86$$

Activity : C

There are two series of index numbers -D for disposable personal income and S for sale of a company. The mean and standard deviation of the D series are 120 and 15 respectively and of the S series 115 and 10. The coefficient of correlation between the two series is +.75. From the given information obtain a linear equation for estimating the values of S for different values of D. How will you interpret the values of S corresponding to different values of D obtained from the equation? Can the same equation be used for estimating the values of D for different values of S? If not, state why not, and give the appropriate equation.

_____

_____

Answer : S = 55 + .5D; NO (21)

12.7 Standard Error of Estimate

Best estimate of dependent variables can be done for the given values of independent variable with the help of regression lines. But to examine the accuracy of such estimates, 'Standard error of estimate' is used. Standard error of estimate is the average of deviations of real values and computed or trend values of dependent series. This is square root of unclear or unexplained deviation. This is calculated as follows:

S. E. X on Y

(i) $$S_{xy} = \sqrt{\frac{\sum (X - X_c)^2}{N}}$$

(ii) $$S_{xy} = \sigma_x \sqrt{1 - r^2}$$

(iii) $$S_{xy} = \sqrt{\frac{\sum X^2 - a\sum X - b\sum XY}{N}}$$

S. E. Y on X

(i) $$S_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N}}$$

(ii) $$S_{yx} = \sigma_y \sqrt{1 - r^2}$$

(iii) $$S_{yx} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{N}}$$

Here,

S_{xy} and S_{yx} = Standard error of the estimate of X on Y and Y on X respectively.
X_c and Y_c = Computed values of X and Y series respectively.

With the help of S.E., r can be calculated as follows : -

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} \text{ or } r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

Interpretation of Standard Error of Estimate :

Standard error of estimate has the same relation with regression line as standard deviation has with mean. If the standard error of estimate is big, it means the points are scattered far away from regression line and computed values are not good estimates of real values. If its value is zero, the points will be on regression line (i.e., perfect correlation between the series); and the estimates are equal to real values. In this way standard error of estimate helps us in understanding that how accurate and representative the regression line is.

Illustration : 9

Calculate standard error of Estimate of regression lines from the data given below :-

X	1	2	3	4	5
Y	2	5	3	8	7

Solution :

Calculation of Regression Equations

X	dx (X - 2)	dx^2	Y	dy (Y -4)	dy^2	dxdy
1	-1	1	2	-2	4	2
2	0	0	5	+1	1	0
3	+1	1	3	-1	1	-1
4	+2	4	8	+4	16	8
5	+3	9	7	+3	9	9
Total	**+5**	**15**		**+5**	**31**	**18**

Means of X and Y

$$\bar{X} = A + \frac{\sum dx}{N} = 2 + \frac{5}{5} = 2 + 1 = 3$$

$$\bar{Y} = A + \frac{\sum dy}{N} = 4 + \frac{5}{5} = 4 + 1 = 5$$

Regression Coefficients

(i) X on Y

$$b_{xy} = \frac{\sum dxdy - \frac{\sum dx \times dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}} = \frac{18 - \frac{5 \times 5}{5}}{31 - \frac{(5)^2}{5}} = \frac{18 - 5}{31 - 5} = \frac{13}{26} = .5$$

$\therefore\ b_{xy} = .5$

(ii) Y on X

$$b_{yx} = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}} = \frac{18 - \frac{5 \times 5}{5}}{15 - \frac{(5)^2}{5}} = \frac{18 - 5}{15 - 5} = \frac{13}{10} = 1.3$$

$\therefore\ b_{yx} = 1.3$

Regression Equations

(a) X on Y

$(X - \bar{X}) = b_{xy} (Y - \bar{Y})$

$(X - 3) = .5 (Y - 5)$

$X = 3 + .5Y - .5 \times 5$

$X = 3 - 2.5 + .5Y$

$X = .5 + .5Y$

(b) Y on X

$(Y - \bar{Y}) = b_{yx} (X - \bar{X})$

$(Y - 5) = 1.3 (X - 3)$

$Y = 5 + 1.3X - 1.3 \times 3$

$Y = 5 - 3.9 + 1.3X$

$Y = 1.1 + 1.3X$

Calculation of Standard Error of Estimate

X	x_c	$(x - x_c)$	$(x - x_c)^2$	Y	y_c	$(y - y_c)$	$(y - y_c)^2$
1	1.5	-.5	.25	2	2.4	-.4	.16
2	3	-1	1	5	3.7	1.3	1.69
3	2	1	1	3	5	-2	4
4	4.5	-.5	.25	8	6.3	1.7	2.89
5	4	1	1	7	7.6	-.6	.36
Total			**3.5**				**9.1**

Note : x_c have been calculated by putting the values of Y in the first equation i.e. X = .5 + .5Y. In the same way y_c have been calculated by putting the values of X in Y = 1.1 + 1.3X equation.

$$\Sigma (x - x_c)^2 = 3.5$$

$$S_{xy} = \sqrt{\frac{\Sigma (x - x_c)^2}{N}}$$

$$= \sqrt{\frac{3.5}{5}} = .836$$

$$\Sigma (y - y_c)^2 = 9.10$$

$$S_{yx} = \sqrt{\frac{\Sigma (y - y_c)^2}{N}}$$

$$= \sqrt{\frac{9.10}{5}} = 1.349$$

12.8 Key Words

Regression : Measurement of average relationship between two or more variables in the form of original units is called regression.

Linear Regression : If a straight line is derived from the marked points in scatter diagram, it will be linear regression.

Regression lines : The lines of best fit drawn to show the mutual relationship between X and Y variables are known as regression lines.

Constant 'a' : Constant 'a' is that point on which regression line touches the 'y' axis. In other words it determines the level of the fitted line.

Constant 'b': It determines the slope of the fitted line. It is also named as regression coefficient.

Regression Equation of X on Y : Equation used to estimate the value of X (dependent variable) from the given value of Y (independent variable), it is X = a + bY

Regression Equation of Y on X : By this equation we can estimate the value of Y (dependent variable) from the given value of X (independent variable), it is Y = a + bX.

Standard Error of Estimate : It is the average of the deviations of actual and computed values.

12.9 Self Assessment Questions

1. Define regression. Why are there two regression lines? Under what conditions can there be only one regression line?

2. Distinguish clearly between 'correlation' and 'regression' concepts used in statistical analysis?

3. What do you understand by dependent and independent variable? How are regression coefficients obtained for these variable?

4. Explain the term 'regression' and its utility in economic analysis. How are regression equations derived? Explain with an example.

5. What is the concept of standard error of estimate in regression? State its utility.

6. Write short notes on :

a) Regression coefficient;

b) Standard error of estimate.

7. a) Estimate the yield when rainfall is 9 inches from the following data :

	Mean	S.D.
Yield of Rice (kg. per unit area)	10	8
Annual Rainfall (inches)	8	2
Correlation coefficient r = .5		(Ans. : 12)

b) The coefficient of correlation between ages of husbands and wives in a community was found to be +.8. The average of husband's age was 25 years and that of wives's age 22 years. Their standard deviations were respectively 4 and 5 years. Find with the help of regression equations :

(i) The expected age of husband when wife's age is 20 years.

(ii) The expected age of wife when husband's age is 25 years. (Ans. : X = 23.72, Y = 22 yrs.)

8. The following data relates to the height (X) and weight (Y) of 1000 executives :

Mean Height ($\bar{X}$) = 68", Mean Weight ($\bar{Y}$) = 160 lbs

Standard deviations (σ_x) of height = 3.6" and that of weight (σ_y) = 25 lbs and r = 0.8

Estimate : -

(i) The height of an executive whose weight is 100 lbs.

(ii) The weight of an executive whose height is 5 ft.

Ans. : (i) 61.07", (ii) 115.6 lbs.

9. From the following data obtain the two regression equations :

X :	6	2	10	4	8
Y :	9	11	5	8	7

(Ans. : X = 16.4 - 1.3Y, Y = 11.9 - 0.65 X)

10. If two lines of regression are given by following equations which of these is the line of regression of Y on X and why?

$4x - 5y + 30 = 0$ and $20x - 9y - 107 = 0$

(Ans. : Y on X = $4X - 5Y + 30 = 0$)

11. Following table gives the ages of husbands and wives for 50 newly married couples. Find the two regression lines. Also estimate (a) the age of husband when wife is 20 and (b) age of wife when husband is 30.

Age of	Age of Husbands		
Wives	20-25	25-30	30-35
16-20	9	14	-
20-24	6	11	3
24-28	-	-	7

(Ans : X = 8.03 + .47Y; Y = 12.08+.72X; Y = 26.48; X = 22.13)

12. Only the following results are available from the records of a partially destroyed laboratory.

Variance of X series = 9

Regression equation : $8x - 10y + 66 = 0$

$40x - 18y = 214$

Find out the following values from the above information :

(i) Mean values of X and Y; (ii) Coefficient of correlation between X & Y; and (iii) Standard deviation of Y. Ans.

($\overline{X} = 13,\ \overline{Y} = 17,\ r = .6,\ \sigma_y = 4$)

13. Find out the regression equation by least square method from the following data :

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Ans. (X=-0.5+1.5Y, Y=0.52+0.64X)

12.10 Reference Books

1. Sharma, J.K., Business Statistics
2. Nagar, K.N., Statistical Methods
3. Agarwal, D.R., Business Statistics
4. Yadav, Jain, Mittal, Statistical Methods
5. Gupta, S.P., Statistical Methods
6. Pinnai & Bhagwati, Statistical Methods

Unit - 13 Analysis of Time Series (Secular Trends)

Structure of Unit:

13.0 Objectives

After completing the unit, you will be able to :

- Assess the importance and usefulness of time series analysis.
- Understand the pattern of historical data.
- Forecasting about future trends.

13.1 Introduction

Today's world is characterised by fast changing business and economic scenario. To cope up with this change, it is necessary to predict the future through our knowledge of past. By evaluating past we want to look into future. One method for evaluating historical data collected on time dimension is **Time series analysis**. The technique of time series analysis lies on the fact that past behaviour of a variable quite often serves as a basis of drawing inference about its future trends.

A **Time series** is a sequence of data points, measured typically at successive times, spaced at uniform time intervals. The time interval may be an hour, a day, a week, a month or a year, depending upon the purpose of data collection. Examples of time series are the daily closing value of the National Stock Exchange index, monthly salary of an employee, yearly production of wheat etc. Time series occur frequently when looking at industrial data for example export/ import over the years, production over the years, demand over the years etc. Thus time series have an important place in Economics, Business and Commerce. **Time series analysis** comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. It is used to detect patterns of change in statistical information over regular intervals of time. These patterns help us in arriving at an estimate for the future so as to cope up with the uncertainty about the future. The assumption underlying time series analysis is that the future will behave as per the past. The objective of time series analysis is to identify the pattern and isolate the influencing effects so as to predict the future.

Time series forecasting is the use of a model to forecast future events based on known past events that is to predict data points before they are measured. An example of time series forecasting in econometrics is predicting the opening price of a stock based on its past performance. Forecasting is an essential tool in strategic planning and decision making process of an organisation. Every organisation is interested in knowing the future values of their key decision variables. Forecasting takes the historical data and projects them into future to predict the occurrence of uncertain events. For example a company has to predict sales so as to produce enough products for customers. Thus forecasting is essential to make reliable and accurate estimates of what will happen in future in the face of uncertainty.

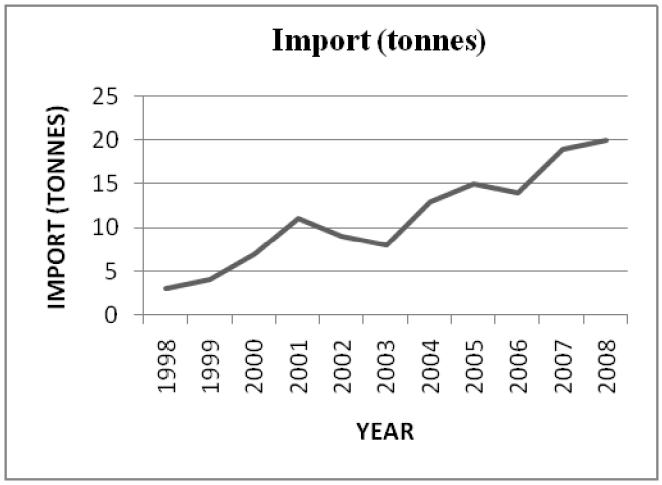
A time series graph is a plot of a variable against time. Analysis of time series starts with plotting observed values of some variable on the vertical axis and time measurement on horizontal axis. Points are plotted against the time series magnitude and joined by means of straight lines. The series graphed in this manner conveys the behaviour of the variable under consideration. A close examination of the graph provides useful insights about the pattern of variations in time series so as to draw meaningful conclusion about the future.

The following table 1.1 and Figure 1.1 presents the import of iron by a company between the years 1998 to 2008.

Table No. 1

Year	Import (tonnes)
1998	3
1999	4
2000	7
2001	11
2002	9
2003	8
2004	13
2005	15
2006	14
2007	19
2008	20

Figure 1: A Time Series Graph



13.2 Applications of Time Series

The usage of time series models is twofold, first is to obtain an understanding of the underlying forces and structure that produced the observed data and second is to fit a model and proceed to forecasting, monitoring or even feedback and feed forward control.

Various applications of Time Series Analysis are as follows:

- Economic Forecasting
- Sales Forecasting
- Budgetary Analysis
- Stock Market Analysis
- Yield Projections
- Process and Quality Control
- Inventory Studies
- Workload Projections
- Utility Studies
- Census Analysis

13.3 Variations in Time Series

There are four types of change or variation involved in time series analysis. These changes are called as components of time series. These are as follows:

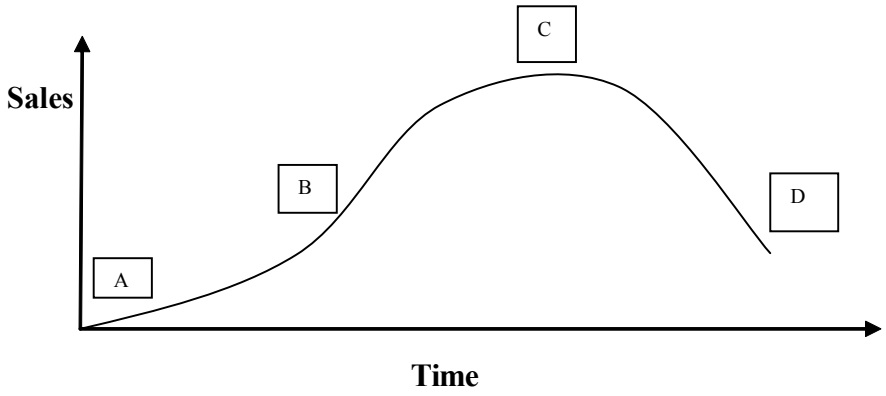
a) Secular trend or Long term trend
b) Cyclic fluctuations
c) Seasonal variations
d) Irregular variation

a) **Secular Trend or Long term trend:** In Secular Trend, the value of variable tends to increase or

decrease over a long period of time. The variable may be population, sales, cost of living, death rates and so on. The trend includes steady movements over a fairly long time and does not include short time oscillations. The trend may be in form of a slow/ fast increase or decrease. These changes occur as a result of general tendency of the data to increase or decrease as a result of some identifiable influences. This trend is most commonly seen in data related to Economics, Business and Population. For example if one considers the death rate of our country, there seems to be a declining trend. This is due to better medical facilities, development in medical sciences, more spending of state in health sector etc. This general trend remains unaffected by major calamities happening in some years.

Trends can be linear or curvilinear. A common example of a curvilinear relationship is 'Product Life Cycle' (PLC). When a new product is introduced, its sales is low (A). As the product gets recognition, the sales grow rapidly (B). After the product is finally established, its sales grow steadily (C). Finally, as the product reaches the end of its life cycle, unit sales begin to decrease (D). A typical PLC can be represented by the following figure:

Figure 2 : A Curvilinear Trend: Product Life Cycle



In practice linear trends are commonly used.

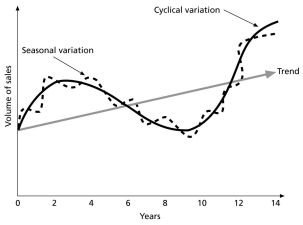
b) **Cyclic fluctuations**: The Cyclic fluctuations refer to regular swings or patterns that repeat over a long period time. Cycles are the upward and downward movements about the trend line over a period of time. The movements are considered cyclic if the fluctuations occur after the time interval of more than one year and it can be as many as 15 to 20 years. In general these cycles are not regular. They may have different periods and intensity (amplitude). The most common example of cyclical fluctuations is business cycle. Business cycles hit a peak above the trend line in some years and hit a low point below the trend line in other years. According to some economists business cycle completes four phases every 12 to 15 years. These four phases are from prosperity, recession, depression and recovery or growth. Although the measurement and prediction of cyclic variation is very important for planning and decision making, the reliability of such measurements remains doubtful, as these cycles do not occur at regular intervals. Also, the cyclic variations are affected by erratic, irregular and random forces which cannot be separated to measure their impact accurately.

c) **Seasonal Variation:** Seasonal variation involves patterns of change within a year that tends to be repeated form year to year. For example, sales of ice cream increases during summers. As these variations are regular pattern, they can be predicted fairly accurately. Some factors that cause seasonal variations are as follows:

i) *Season and climate*: Changes in climate and weather affect the sales of many products. For example, sale of coolers and AC's is always more during summer months. Similarly there is a greater demand for woolen clothes during winters, rain coats in rainy season etc.

ii) *Customs and festivals*: Traditions and customs also affect the patterns of sales. For example sales of paints and furnishing goods is more during Diwali, sales of sweets is more during festivals etc.

Figure 3 : Graph showing Secular trend, Seasonal and Cyclic variations



d) **Irregular Variations:** Irregular Variations are rapid, unpredictable and random changes in the data. The changes are caused by short term unanticipated and non recurring factors/ incidents like wars, strikes, floods, famine, earth quakes etc. The effect of Iraq war and terrorist attacks on crude oil prices is an example of irregular variation.

As the changes are random, it is quite difficult to isolate and measure the value and impact of these erratic movements on forecasting models. The best that can be done is to obtain their rough estimates from past experiences and accordingly make provisions about such abnormalities during normal times in business.

13.4 Decomposition of Time Series

Analysis of Time Series consists of (i) identification of various factors which cause the variation in the time series and (ii) separating, analysing and measuring the effect of these factors independently.

The objective of decomposition of time series is to break time series into four components viz., Trend (T), Cyclic (C), Seasonal (S) and Irregular (I), so as to isolate influence of each of the four components on the actual series.

The two common models used for decomposition of a time series are:

(i) Multiplicative model and (ii) Additive Model.

(i) **Multiplicative model** : According to multiplicative model, the value of time series (Y) can be found by multiplying four components. The effect of four components on the time series is interdependent. The model can be expressed as:

$$\mathbf{Y = T \times C \times S \times I}$$

This model is more common than additive model. It is appropriate in situations where the effect of four components is measured in relative terms rather than absolute values.

(ii) **Additive Model :** This model is based on the assumption that the effect of four components can be estimated by adding the various components of a time series. It is expressed as follows:

$$\mathbf{Y = T + C + S + I}$$

This model is less common. It is appropriate in situations when variations in the time series are in absolute values and can be separated to each of the four parts. Here each part is independent. Each part can be negative or positive.

Activity A

Answer the following questions:

1. Explain the purpose for which the time series analysis is applied to data collected over a period of time.
2. What do you understand by time series analysis?
3. What are the different components of time series?
4. Explain briefly the additive and multiplicative models of time series.
5. What are the four main components of a time series?
6. Differentiate between seasonal and cyclical variation.

7. What do you understand by time series forecasting?
8. Which of the four components of time series might the Ministry of Agriculture use to describe a 9 year weather pattern?
9. How would a war be described in a time series?
10. What are the various components of a time series? Explain the general growth and decline of the steel industry over last two decades?

13.5 Measurement of Secular Trend

Secular trend measures the long term direction of series. The study of secular trends helps in:

a) identification of a historical pattern.
b) projecting past patterns into future.
c) Eliminating the trend components from the series.

The most important methods for measuring Secular Trend are as follows:

i) Free Hand Curve method
ii) Semi Averages Method
iii) Moving Averages Method
iv) Method of Least Square

i) **Free Hand Curve method**

It is the simplest method. It consists of plotting the time series values on a graph paper and then drawing a free hand smooth curve through these points so as to reflect the long term tendency of the data. This method is based on inspection, hence is subjective in nature that is only a specialist with long experience who knows the behaviour of the variable can draw a line by mere inspection.

Illustration 1:

Year	1998	1999	2000	2001	2002	2003	2004	2005
Sales (000's Rs)	84	94	96	87	98	103	96	108

Solution:

ii) **Semi Averages Method :** In semi average method, the data is divided into two equal parts and arithmetic means are calculated for each of the parts. These are called semi averages. These two points are plotted corresponding to the midpoint of the respective class intervals covered by each part. The line joining these two points gives the straight trend line. This line can be extended upward and downward to get an estimate of future and past values. If the number of years is odd then we leave the middle year. For example if the number of years is 9, then we leave the middle year i.e. 5^{th} year and the remaining years are divided into two equal parts of 4 years each.

Illustration 2:

Fit a trend line with the help of semi averages method to the following data:

Year	1998	1999	2000	2001	2002	2003
Sales (000's Rs)	12	14	18	10	8	16
Year	2004	2005	2006	2007	2008	2009
Sales (000's Rs)	17	12	16	22	15	20

The given time series will be divided into two equal parts of 6 years each and mean of each part is calculated.

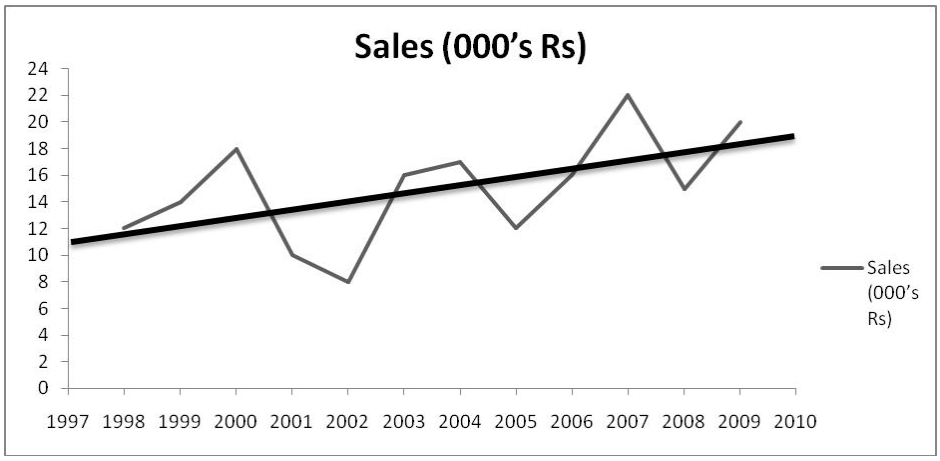
$$\text{Part I: } \overline{x}_1 = \frac{12 + 14 + 18 + 10 + 8 + 16}{6} = 13$$

Part II: $\overline{x_2} = \frac{17 + 12 + 16 + 22 + 15 + 20}{6} = 17$

The arithmetic mean of Part I will be plotted between the middle points of the time series i.e. between 2000-2001. Similarly arithmetic mean of part II will be plotted between the years 2007-2008 on a graph paper.

Year	Sales in '000 Rs	Semi average
1998	*12*	
1999	14	
2000	18	13
2001	10	
2002	8	
2003	16	
Year	Sales in '000 Rs	Semi average
2004	17	
2005	12	
2006	16	17
2007	22	
2008	15	
2009	20	

Figure 4



Advantages:

a. It is not based on personal judgment; hence, same trend line will be obtained every time.
b. Easy to understand
c. Trend line can be extended upward and downward to estimate future and past values.

Disadvantages:

a. It is based on the assumption of linear trend.
b. Use of arithmetic mean: arithmetic mean has its own limitations as it may not be appropriate method in some cases.

Moving Average Method

The moving average method is based on the concept that if the observations at a point in time are averaged out with the values immediately before and after it, it will smooth out the irregular component of time series. This is a very common method to determine secular trend. This method is subjective as it depends on selection of the length of period for calculating moving average for example 3 yearly moving average, 4 yearly moving average, 5 yearly moving average etc. So as to remove the effect of cyclical variations, the length of the period chosen for moving average should be an integer value that corresponds to the estimated average length of a cycle in the series.

If we want to compute a 3- year moving average from a time series having in total nine years, the first three values will be added up pertaining to first three years and its average is calculated. Similarly, the second moving average will be computed by taking the average of next three years, leaving the first year.

Illustration 3:

Fit a trend line with the help of moving averages method to the following data:

Month	Jan	Feb	Mar	Apr	May	Jun
Sales (000's Rs)	21	20	21	25	26	22
Month	Jul	Aug	Sep	Oct	Nov	Dec
Sales (000's Rs)	23	24	30	31	33	29

Solution:

The moving average will be calculated for the months in a group of three viz., Jan., Feb. and Mar, then Feb, Mar and Apr. Mar, Apr and June and so on.

Part I: $\overline{x_1} = \frac{21 + 20 + 21}{3} = 20.66$

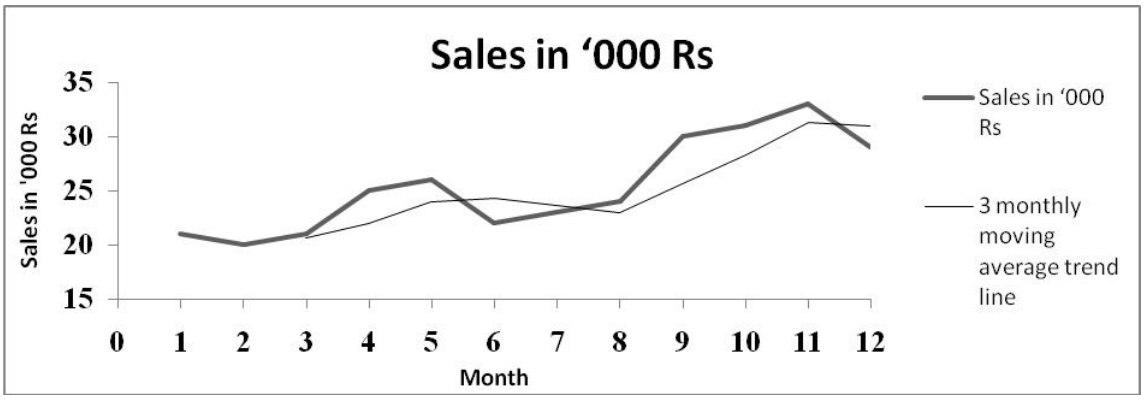
Part II: $\overline{x_2} = \frac{20 + 21 + 25}{3} = 22$

Part III: $\overline{x_3} = \frac{21 + 25 + 26}{3} = 24$, and so on.

From the table below, we can see that the first moving average value of 20.66 will be placed against the month of April. Similarly the next moving average value of 22 will be placed against May.

Months	Sales in '000 Rs	3 year moving average	Fits
Jan	21		----------
Feb	20	20.66	----------
Mar	21	22	----------
Apr	25	24	20.66
May	26	24.33	22
Jun	22	23.66	24
Jul	23	23	24.33
Aug	24	25.66	23.66
Sep	30	28.33	23
Oct	31	31.33	25.66
Nov	33	31	28.33
Dec	29		31.33

Figure 5 : 3 month moving average plot



If we want to draw a 4 year moving average trend line for the data given in table 2, it will be computed as follows. First we will calculate four yearly moving

Months	Sales in '000 Rs	4 monthly total	Moving total of pairs	Four monthly moving average
Jan	21			
Feb	20			
Mar	21	87		22.38
Apr	25	92	179	23.25
May	26	94	186	23.75
Jun	22	96	190	23.88
Jul	23	95	191	21.75
Aug	24	79	174	23.38
Sep	30	108	187	28.25
Oct	31	118	226	30.13
Nov	33	123	241	----
Dec	29	----	----	
		----		

iii) **Method of Least Square**

This is the method in which a trend line is determined using a mathematical equation of either a straight line or a parabolic curve. As this method is based on mathematical equation it is not subjective and hence is more accurate. The equation used for fitting a straight line is : $\mathbf{Y_c = a + bX,}$ where 'Y_c' is the predicted value of dependent variable, 'X' is independent variable which in this case is *time*. Further, 'a' is intercept and 'b' is slope. Both 'a' and 'b' are constant.

The trend line has the following properties:

i) The summation of all vertical deviations from it is zero, that is $\sum (Y - Y_c) = 0$.

ii) The sum of all vertical deviations squared is minimum that is $\sum (Y - Y_c)$ is least.

iii) The line goes through mean values of X and Y

Fitting a straight line

The equation used for fitting a straight line is : $\mathbf{Y_c = a + bX,}$ where

'Y_c' is the predicted value of dependent variable,

'X' is independent variable i.e.*time*.

'a' is intercept and

'b' is slope. (Both 'a' and 'b' are constant)

For solving 'a' and 'b', the following two normal equations will be solved

$\mathbf{\sum Y = N.a + b \sum X}$

$\mathbf{\sum XY = a \sum X + b \sum X^2}$

The independent variable 'X' is given order number 1,2,3,4,5 … for calculating $\mathbf{\sum X.}$

Illustration 4:

The following table represents the data for production of steel by a company. Fit a straight trend and estimate the profit for the year 2007

Year	2000	2001	2002	2003	2004	2005	2006
Production (000's Kgs)	400	800	700	900	1000	800	1100

Solution

Year	Production (000's Kgs)	X	XY	X^2	Trend Values Y_c= 471.43 + 85.71X
2000	400	1	400	1	557.14
2001	800	2	1600	4	642.85
2002	700	3	2100	9	728.56
2003	900	4	3600	16	814.27
2004	1000	5	5000	25	899.98
2005	800	6	4800	36	985.69
2006	1100	7	7700	49	1071.40
N=7	∑Y= 5700	∑X=28	∑XY= 25200	$\sum X^2$= 140	$\sum Y_c$ = 5700

The normal equations are:

$\sum Y = N.a + b \sum X$ i.e. 5700 = 7a + 28b (equation 1)

$\sum XY = a \sum X + b \sum X^2$ i.e 25200 = 28a + 140b (equation 2)

Multiplying equation (1) by 4 we get : 22800 = 28a + 112b (equation 3)

On subtracting eq. (3) from eq. (2) we get : 2400 = 28 b

$$b = \frac{2400}{28} = 85.71 \qquad a = \frac{3300}{7} = 471.43$$

So that

Thus 7a + 28 (85.71) b = 5700 or 7a + 2400 = 5700 or

The trend line is :

$Y_c = a + bX$ i.e. $Y_c = 471.43 + 85.71X$

To get the trend value for the year 2007, the value of X will be 8, so that:

$Y_c = 471.43 + 85.71$ (8)

Y_c = 1157.11 9 (000' Kgs)

Thus the estimated production for the year 2007 will be 11,57,110 Kgs.

13.6 Summary

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. It is used to detect patterns of change in statistical information over regular intervals of time. There are four types of change or variation involved in time series analysis. These changes are called as components of time series. These are: Secular trend or Long term trend, Cyclic fluctuations, Seasonal variations and Irregular variation. The two common models used for decomposition of a time series are: Multiplicative model and Additive Model. Secular trend measures the long term direction of series. The study of secular trends helps in identification of a historical pattern, projecting past patterns into future and eliminating the trend components from the series. The most important methods for measuring Secular trend are as follows: Free hand curve method, Semi averages method, Moving averages method and Method of least square. As Method of least square is based on mathematical equation, it is not subjective and hence is more accurate.

13.7 Key Words

Time Series : It is a sequence of data points, measured typically at successive times spaced at uniform time intervals.

Time Series Forecasting : It is the use of a model to forecast future events based on known past events: to predict data points before they are measured.

Secular Trend : In it the value of variable tends to increase or decrease over a long period of time.

Cyclic Fluctuations : It refer to regular swings or patterns that repeat over a long period time.

Seasonal Variation : It involves patterns of change within a year that tends to be repeated form year to year.

Irregular Variations : These are rapid, unpredictable and random changes in the data.

Method of Least Square : This is the method in which a trend line is determined using a mathematical equation of either a straight line or a parabolic curve.

13.8 Self Assessment Questions

1 Fit a trend line to the following data by free hand method:

Year	2001	2002	2003	2004	2005	2006	2007	2008
Production (000's units)	80	82	88	85	87	92	90	94

2 Fit a trend line to the following data by semi average method:

Year	2000	2001	2002	2003	2004	2005	2006
Sales (000's Rs.)	200	212	248	232	224	256	240

3 Fit a trend line with the help of 3 year moving averages method to the following data :

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Sales (000's Rs)	15	17	19	22	21	20	18	22	23	27

(Answer: 17, 19.33, 20.67, 21, 19.67, 20, 21, 24).

4 Fit a trend line with the help of 6 year moving averages method to the following data :

Year	Profit (000's Rs.)
1995	50
1996	80
1997	110
1998	100
1999	90
2000	75
2001	80
2002	110
2003	150
2004	200
2005	160
2006	130
2007	150

(Answer : 86.67, 91.67, 97.50, 109.17, 123.33, 133.75, 144.17)

5 Fit a trend line to the following data of sales of a company by using least square method and find the trend values:

Year	2000	2001	2002	2003	2004	2005	2006	2007
Sales (in units)	482	428	385	360	329	290	270	232

(Answer: $Y_c = 500 - 34\,X$)

13.9 Reference Books

1. Gupta, S.P.. Statistics, Sultan Chand & Sons.
2. Sharma, J K, Business Statistics, Pearson Education.
3. Chandan, J S, Business Statistics, Vikas Publishing House Pvt Ltd, New Delhi.
4. Hooda R P, Statistics for Business and Economic, Macmillan, New Delhi.
5. Levin and Rubin. Statistics for management, Prentice Hall of India Ltd., New Delhi.

Unit - 14 Analysis of Time Series (Seasonal, Cyclical and Irregular Variations)

Structure of Unit:

14.0 Objectives

After studying this unit you would be able to:

- Understand the meaning of short term variations.
- Measure the seasonal variations.
- Understand the merits and limitations of different methods.
- Calculate cyclical variations.
- Understand the impact of irregular fluctuations.

14.1 Introduction

As we have already discussed that a Time Series is affected by long term and short term fluctuations collectively. If, from the original data long term variations (as calculated in previous chapter) removed, the balance will be short term fluctuations.

Thus, Short Term Fluctuations = Original Data - Trend values.

These short term fluctuations can be of regular or irregular nature. As per **Croxton & Cowden** : "A periodic movement is one which occurs, with some degree of regularity, within a definite period". These regular fluctuations can be divided into two parts (a) Seasonal variations (b) Cyclical variations. Irregular or random variations are which do not repeat in definite pattern. According to Patterson- "The irregular variations in a time series is composed of non-recurring, sporadic forces which are not described as or attributed to trend, cyclical or seasonal factors." In this unit we will study the methods of measuring these short term variations.

14.2 Measurement of Seasonal Variations

Seasonal variations have been defined as repetitive and predictable movements around the trend line in a period of one year or less. The word season includes any kind of variation (like climate and weather conditions, custom, traditions habits etc.) which is of periodic nature and whose repeating cycles are of

relatively short duration. Due to the predictable nature, we can plan in advance to meet these variations. In order to calculate seasonal variations, we first eliminate as far as possible the effects of trend, cyclical variations and irregular variations. Seasonal variations are measured in Indices form, so these are also called Seasonal Indices.

These indices may be used for economic forecasting and managerial control. Seasonal patterns that directly influence company's sales, purchase, production, inventory or employment policies are examined, which takes management in better control position.

The following methods are used for measurement of seasonal variations:

1. Simple Average Method or Seasonal Variation Index.
2. Seasonal Variation through Moving Averages Method.
3. Chain or Link Relative Method.
4. Ratio to Moving Averages Method.
5. Ratio to Trend Method.

14.2.1 Simple Average Method or Seasonal Variation Index Method

This is the simplest method of isolating seasonal variations. The basic assumption of this method is that the series contains only the seasonal and irregular fluctuations.

The following steps are necessary for calculating seasonal Indices:

i. Arrange the unadjusted data either on yearly, quarterly or monthly basis.
ii. Find the totals on the basis of quarters or months.
iii. Calculate quarterly or monthly average by dividing the total by number of years.
iv. Find out general average by dividing this total by 12.
v. Taking the average of monthly averages as 100, calculate seasonal indices as follows -

$$\text{Seasonal Index} = \frac{\text{Average value of the month or quarter}}{\text{General Average}} \text{ X } 100$$

Illustration: 1

Calculate seasonal Indices by simple average method from the following data:

Year	2005	2006	2007	2008	2009
Summer	31	35	43	54	67
Monsoon	60	54	59	53	59
Autumn	65	68	99	87	102
Winter	91	71	86	73	65

Solution:

Computation of Seasonal Indices

Year	Sumer	Monsoon	Autumn	Winter
2005	31	60	65	91
2006	35	54	68	71
2007	43	59	99	86
2008	54	53	87	73
2009	67	59	102	65
Total	230	285	421	386
Average	46	57	84.2	77.2
Seasonal Indices	69.6	86.23	127.38	116.8

$$\text{General Average} = \frac{\text{Total Seasonal Averages}}{\text{No.of Seasons}} = \frac{46 + 57 + 84.2 + 77.2}{4}$$

$$= \frac{264.4}{4} = 66.1$$

$$\text{Seasonal Indices} = \frac{\text{Seasonal Average}}{\text{General Average}} \text{ x } 100$$

$$\text{Seasonal Index for Summer} = \frac{46}{66.1} \text{ x } 100 = 69.6, \qquad \text{for Monsoon} = \frac{57}{66.1} \text{ x } 100 = 86.23$$

$$\text{for Autumn} = \frac{84.2}{66.1} \text{ x } 100 = 127.38, \qquad \text{and for winter} = \frac{77.2}{66.1} \text{ x } 100 = 116.8$$

Illustration: 2

Calculate Seasonal Indices from the following data:

Months	2008	2009	2010
January	60	63	45
February	70	68	27
March	60	50	40
April	52	38	30
May	44	36	40
June	34	26	36
July	22	18	26
August	37	31	31
September	42	31	23
October	31	27	29
November	45	42	42
December	50	42	25

Solution:

Computation of Seasonal Indices

Months	2008	2009	2010	Total (3 yearly)	Monthly Average	Seasonal Indices
Jan.	60	63	45	168	56	142.67
Feb.	70	68	27	165	55	140.13
March	60	50	40	150	50	127.39
April	52	38	30	120	40	101.91
May	44	36	40	120	40	101.91
June	34	26	36	96	32	81.53
July	22	18	26	66	22	56.05
Aug.	37	31	31	99	33	84.07
Sept.	42	31	23	96	32	81.53
Oct.	31	27	29	87	29	73.88
Nov.	45	42	42	129	43	109.55
Dec.	50	42	25	117	39	99.36

$$\text{General Average} = \frac{\text{Total of Monthly Averages}}{\text{No. of Months}} = \frac{471}{12} \text{ or } 39.25$$

$$\text{Seasonal Indices} = \frac{\text{Monthly Averages}}{\text{General Average}} \text{ x } 100$$

Like for the month of January = $\frac{56}{39.25}$ x 100 = 142.67

Similarly for other months.

Illustration: 3

Calculate seasonal Indices by simple average method from the following data :-

Year	I	II	III	IV
2002	3.7	4.1	3.3	3.5
2003	3.7	3.9	3.6	3.6
2004	4.0	4.1	3.3	3.1
2005	3.3	4.4	4.0	4.0

Solution

Calculation of Seasonal Indices

Year	I	II	III	IV
2002	3.7	4.1	3.3	3.5
2003	3.7	3.9	3.6	3.6
2004	4.0	4.1	3.3	3.1
2005	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.2	14.2
Average	3.675	4.125	3.550	3.550
Seasonal Indices	98.7	110.8	95.3	95.3

General Average = $\frac{3.675 + 4.125 + 3.550 + 3.550}{4} = 3.725$

Seasonal Index = $\frac{\text{Quarterly Average}}{\text{General Average}}$ x 100

Seasonal Index for I quarter = $\frac{3.675}{3.725}$x100 = 98.7 and so on

Merits and Limitations of this Method:

This is the simplest of all the methods. But it assumes that there is no trend component, which is not justified. Most of the economic series have trends. Secondly the effect of cycles on the original value cannot be fully eliminated by averaging process. Thus, this method is useful only where no definite trend exists.

14.2.2 Moving Average Method

This method is based on additive model. Except cyclical variations all other variations like trend, seasonal variations and irregular fluctuations are determined by it.

The following procedure is adopted:

i. Find moving averages of data. In the case of monthly data, twelve monthly moving average will be calculated. If the data are given on quarterly basis, four quarterly moving average will be calculated.

ii. Find short term fluctuations = original data - moving average or S+C+I = O-T

iii. Average these short term fluctuations in separate table.

iv. Calculate seasonal variations by averaging the monthly or quarterly data.

Illustration: 4

Calculate seasonal fluctuations by moving average method from the following data:

Year	I	II	III	IV
2001	70	90	80	100
2002	65	60	90	95
2003	80	85	90	80
2004	85	70	85	90

Solution:

Computation of Seasonal Variations

Years	Quarters	Values (O)	Quarterly Total	Total Centered	Moving Averages (T)	Short term Oscillations (O-T)	Seasonal Variations
2001	I	70	-	-	-	-	-
	II	90	-	-	-	-	-
			340				
	III	80		675	84.4	-4.4	3.93
			335				
	IV	100		640	80	20	9.16
			305				
2002	I	65		620	77.5	-12.5	-5.2
			315				
	II	60		625	78.125	-18.125	-9.99
			310				
	III	90		635	79.4	10.6	3.93
			325				
	IV	95		675	84.4	10.6	9.16
			350				
2003	I	80		700	87.5	-7.5	-5.2
			350				
	II	85		685	85.6	-0.6	-9.99
			335				
	III	90		675	84.4	5.6	3.93
			340				
	IV	80		665	83.125	-3.125	9.16
			325				
2004	I	85		645	80.6	4.4	5.2
			320				
	II	70		650	81.25	-11.25	-9.99
			330				
	III	85	-	-	-	-	-
	IV	90	-	-	-	-	-

Calculation of Average Seasonal Variations

Years	I Quarter	II Quarter	III Quarter	IV Quarter
2001	-	-	-4.4	20
2002	-12.5	-18.125	10.6	10.6
2003	-7.5	-0.6	5.6	-3.125
2004	4.4	-11.25	-	-
Total	-15.6	-29.975	11.8	27.475
Average	-5.2	-9.99	3.93	9.16

Merits and Limitations:

Moving average method is also a simple and easy to understand method. This method is suitable if values have a particular cycle. Where the cycle is not definite, this method will not give correct results.

14.2.3 Ratio to Moving Average Method

This is the most widely used method of measuring seasonal variations. This is based on multiplicative model. The following procedure is adopted for determining seasonal variations -

1. Find monthly or quarterly moving averages of the original data.

2. Express the original data for each month as a percentage of the corresponding moving average. This is called moving average ratio.

$$\text{Moving Average Ratio} = \frac{\text{Original Data}}{\text{Moving Average}} \text{x}100 \text{ or } = \frac{\text{O}}{\text{T}} \text{x}100$$

3. Arrange these moving average ratios in a separate table according to months or quarters as the case may be. Find their average.

4. Calculate General average from the following formula -

$$\text{General Average} = \frac{\text{Sum of averages of moving average ratios}}{\text{No. of months or Quarters}}$$

5. $$\text{Seasonal Indices} = \frac{\text{Quarterly or Monthly Average}}{\text{General Average}} \text{ x } 100$$

Illustration: 5

From the data given below, calculate the Seasonal Indices through ratio to moving average method:

Years	I Quarter	2nd Quarter	3rd Quarter	4th Quarter
2003	35	40	38	42
2004	40	38	37	39
2005	35	41	42	38
2006	36	45	44	38

Solution:

Computation of Ratio to Moving Averages

Years	Quarters	Given Values	4 Figures Moving Total	Paired Values	4 Yearly Moving Average	Ratio to Moving Average
2003	I	35	-	-	-	-
	II	40	-	-	-	-
			155			
	III	38		315	39.4	96.45
			160			
	IV	42		318	39.75	105.66
			158			
2004	I	40		315	39.4	101.52
			157			
	II	38		311	38.87	97.76
			154			
	III	37		303	37.88	97.68
			149			
	IV	39		301	37.63	103.64
			152			
2005	I	35		309	38.63	90.60
			157			
	II	41		313	39.12	104.81
			156			
	III	42		313	39.12	107.36
			157			
	IV	38		318	39.75	95.60
			161			
2006	I	36		324	40.5	88.89
			163			
	II	45		326	40.75	110.43
			163			
	III	44	-	-	-	-
	IV	38	-	-	-	-

Calculation of Seasonal Indices

Years	I	II	III	IV
2003	-	-	96.45	105.66
2004	101.52	97.76	97.68	103.64
2005	90.60	104.81	107.36	95.60
2006	88.89	110.43	-	-
Total	281.01	313	301.49	304.9
Average	93.67	104.33	100.5	101.63
Seasonal Indices	93.64	104.3	100.5	101.6

$$\text{General Index} = \frac{\text{Total of Quarterly Average}}{\text{No. of Quarters}}$$

$$= \frac{93.67 + 104.33 + 100.5 + 101.63}{4} = 100.03$$

$$\text{Seasonal Index} = \frac{\text{Quarterly Average}}{\text{General Average}} \text{x } 100$$

Seasonal Index for I Quarter $= \frac{93.67}{100.03} \text{ x } 100 = 93.64$ and so on

Merits and Limitations:

Though this is more widely used method in practice than other methods, yet one drawback of this method is that seasonal indices cannot be obtained for each month or quarter for which data are available. If data is given on monthly basis six months in the beginning and six months in the and are left out. Similarly if data are in quarterly form two quarters in the beginning and two in the end are left out for which we cannot calculate seasonal indices.

14.2.4 Ratio to Trend Method

This method is based on multiplicative model. It assumes that seasonal variations for a given period is constant fraction of trend. So trend is eliminated when the ratios are computed. Irregular variations are supposed to disappear when the ratios are averaged. The procedure of computation of seasonal indices is as follows:-

i. Trend values are obtained by the method of Least Squares.
ii. Divide the original data by the corresponding trend values and multiply these ratios by 100. It is ratio to trend.

$$\text{Ratio to Trend} = \frac{\text{Original Data}}{\text{Trend Value}} \text{ x } 100$$

iii. The values so obtained are now free from trend. The average of all these ratios to trend will be seasonal variations. This is also called general average.

iv. The seasonal Index for each month is expressed as a percentage of the average of seasonal average i.e.

$$\text{Seasonal Index} = \frac{\text{Seasonal Average}}{\text{Average of Seasonal Average or (General Average)}} \text{ x } 100$$

This gives the final seasonal Index.

Illustration: 6

Calculate seasonal Indices by ratio to trend method from following data:-

Year	1st Qtr.	2nd Qtr.	3rd Qtr.	4th Qtr.
2002	36	34	30	40
2003	36	48	52	44
2004	44	54	50	52
2005	56	74	70	60
2006	80	90	90	80

Solution:

Calculation of Yearly Trend Values

Year	Yearly Total	Average (y)	Origin 2004 (x)	(xy)	x^2	Yearly Trend Values Yc=56 + 12x
2002	140	35	-2	-70	4	56 + 12 x-2 = 32
2003	180	45	-1	-45	1	56 + 12 x-1 = 44
2004	200	50	0	0	0	56 + 12 x 0 = 56
2005	260	65	1	65	1	56 + 12 x 1 = 68
2006	340	85	2	170	4	56 + 12 x 2 = 80
N = 5	1120	280	0	120	10	280

Trend equation = Y = a + bx

$$a = \frac{\Sigma y}{N} = \frac{280}{5} = 56 \qquad b = \frac{\Sigma xy}{N} = \frac{120}{10} = 12$$

Thus the annual increment is of 12 (value of b), so quarterly increment is of $= \frac{12}{4} = 3$

Calculation of quarterly trend values:

Trend value for 2002 is 32 and quarterly increase rate is 3 which must come between 2nd and 3rd Quarter, hence trend value for 2nd Quarter will be $32 - \frac{3}{2} = 30.5$ and trend value for 3rd Quarter will be $32 + \frac{3}{2} = 33.5$, trend value for 1st and 4th Quarter will be 30.5 - 3 = 27.5 and 33.5 + 3 = 36.5 respectively. In this way, we will get the trend values of all the quarters of all the years as follows:

Computation of Quarterly Trend Values

Year	1st Qr.	2nd Qr.	3rd Qr.	4th Qr.	Total
2002	27.5	30.5	33.5	36.5	128.0
2003	39.5	42.5	45.5	48.5	176.0
2004	51.5	54.5	57.5	60.5	224.0
2005	63.5	66.5	69.5	72.5	272.0
2006	75.5	78.5	81.5	84.5	320.0
					1120

Calculation of Ratio to Trend and Seasonal Indices

Year	Ist Qr.	2nd Qr.	3rd Qr.	4th Qr.
2002	130.9	111.5	89.6	109.6
2003	91.1	113	114.3	90.7
2004	85.4	99.1	86.9	85.9
2005	88.2	111.3	100.7	82.8
2006	106	114.6	110.4	94.7
Total	501.6	549.5	501.9	463.7
Average	100.32	109.92	100.38	92.74
Seasonal Indices	99.49	108.99	99.55	91.97

General Average $= \frac{100.32 + 109.9 + 100.38 + 92.74}{4} = 100.835$

Seasonal Indices $= \frac{\text{Quarterly Average}}{\text{General Average}} \times 100$

Merits and Limitations:

This method is simple to compute and easy to understand. This method is certainly a more logical procedure for measuring seasonal fluctuations. There is no loss of data as occurs in the case of moving averages. The main drawback of this method is that if there are pronounced cyclical fluctuations in the series, the trend can never follow the actual data as closely as a twelve month moving average does.

14.2.5 Chain or Link Relative Method

This method involves the following steps:

i. Calculate the link relatives of each month or quarterly figures by following formula:

$$\text{Link Relative} = \frac{\text{Value of Current Season}}{\text{Value of Previous Season}} \text{x} 100$$

ii. Calculate Average of Link Relatives for each season. $= \frac{\text{(Total of LR of a Season)}}{\text{No. of Seasons}}$

For first season no. of seasons will be one less than the other seasons.

iii. Convert these averages into chain relatives, which is calculated as follows:

$$= \frac{\text{Link relative for current quarter x Chained index for previous quarter}}{100}$$

Chain index for 1st quarter will have to be assumed as 100.

iv. Calculate the chain relative of the first quarter on the basis of last quarter.

v. Chained relative computed in point (iv) may differ from base i.e. it should be equal to 100. This is due to long term secular trend. In this situation a correction is done in the chain relatives. Calculate the difference between first quarter's chain relative computed in point (iv) and 100. The difference is divided by the number of seasons. This resulting factor (called correction factor) is multiplied by 1,2,3,4 (and so on) deducted respectively from the chain relatives of the 2nd, 3rd, 4th (and so on.....) seasons. These are corrected chain relatives. If the difference is negative it is added to the chain relatives.

vi. Find the general averages of these corrected chain relatives taking it as base (100). Change the corrected chain relatives in the percentage form. It provides the required seasonal indices. Sum of seasonal indices must be equal to 100 x number of seasons.

The following example will illustrate the process:

Illustration 7: Find out seasonal variation indices by the link relatives method from the following data:

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1997	45	54	72	60
1998	48	56	63	56
1999	49	63	70	65
2000	52	65	75	72
2001	60	70	84	77

Solution:

Calculation of Quarterly Seasonal Indices by the Method of Link Relatives

Year	Link relatives			
	I Quarter	II Quarter	III Quarter	IV Quarter
1997	-	120	133	83
1998	80	117	113	89
1999	88	129	111	92
2000	80	125	115	96
2001	83	117	120	79
Total L.R.	331	608	592	339
Average of L.R.	82.8	121.6	118.4	87.8
Chain Relatives	100	121.6	144	126.5
Corrected Chain Relatives	100	120.4	141.6	122.9
Seasonal Indices	82.5	99.4	116.8	101.3

Calculations have been made in the following way -

i. Link Relatives $= \frac{\text{Value of current season or quarter}}{\text{Value of previous season or quarter}} \times 100$

Link Relative for II Quarter of First year $= \frac{54}{45} \times 100 = 120$

III Quarter $= \frac{72}{54} \times 100 = 133$ and so on ... ...

ii. Chain Relatives for I Quarter has been assumed to be 100

$$\text{Chain Relatives} = \frac{\text{L. R. of current quarter x Chained Index for previous quarter}}{100}$$

$$\text{Chain Relatives for II Quarter} = \frac{121.6 \times 100}{100} = 121.6$$

$$\text{Chain Relatives for III Quarter} = \frac{118.4 \times 121.6}{100} = 144.0$$

$$\text{Chain Relatives for IV Quarter} = \frac{87.8 \times 144}{100} = 126.5$$

$$\text{Chain Relatives for I Quarter} = \frac{82.8 \times 126.5}{100} = 104.7$$

Difference due to presence of trend = 104.7 - 100 = 4.7

$$\text{Difference per Quarter} = \frac{4.7}{4} = 1.2 \text{ approx.}$$

iii. Adjusted Chain Relatives: I Quarter = 100

II Quarter = 121.6 - (1.2 x 1) = 121.6 - 1.2 = 120.4

III Quarter = 144 - (1.2 x 2) = 144.0 - 2.4 = 141.6

IV Quarter = 126.5 - (1.2 x 3) = 126.5 - 3.6 = 122.9

iv. General Average $= \frac{100+120.4+141.6+122.9}{4} = 121.25$

v. Seasonal Indices = $\frac{\text{Corrected Chain Relatives}}{\text{General Average}}$ X 100

I Quarter = $\frac{100}{121.25}$ x 100 = 82.5 III Quarter = $\frac{141.6}{121.25}$ x 100 = 116.8

II Quarter = $\frac{120.4}{121.25}$ x 100 = 99.4 IV Quarter = $\frac{122.9}{121.25}$ x 100 = 101.3

Link relatives method is the most difficult method of calculating seasonal fluctuations. Otherwise it gives better results.

Which method to use:

Selection of method depends upon the nature of data and object of investigation. No one method is suitable for all the cases. If trend does not affect seasonal fluctuations then additive method may be used. Simple average method is appropriate where we do not find trend and cycle. In general, it may be said that because of theoretical and practical advantages, ratio to moving average method should be preferred to other methods.

Illustration: **8**

The seasonal indices of the sale of readymade garments of a particular type in a certain store are given below:

Quarter	Seasonal Index
I Jan - March	98
II April - June	90
III July – Sept	82
IV Oct. - Dec.	130

If the total sales in the first quarter of a year be worth Rs. 20,000, determine how much worth of garments of this type should be kept in stock by the store to meet the demand in each of the remaining quarters.

Solution:

Calculation of Estimated Stock of Readymade Garments

Quarter	Seasonal Index	Estimated stock (Rs.)
Jan. - March	98	20000
April - June	90	$20{,}000\text{x}\frac{90}{98} = 18367$
July - Sept.	82	$20{,}000\text{x}\frac{82}{98} = 16735$
Oct. - Dec.	130	$20{,}000\text{x}\frac{130}{98} = 26531$

14.3 Measurement of Cyclical Variations

Cyclical variations are long term, recurrent up and down movements around secular trend. Their period of time is longer than that of seasonal variations. They take place because of business cycle. According to Burns and Mitchell - "Successive waves of expansion and contraction that occur at about the same time in many economic activities is known as business or trade cycle." Trade cycle move through four phases,

namely (i) Prosperity (ii) Recession (iii) Depression and (iv) Recovery. These phases have different periods and amplitudes but they are constantly repeated in the order given as a cycle completes its swing.

The study of cyclical fluctuations becomes useful in framing suitable policies to control the extreme effects of the cycle. But they are the most difficult type of economic fluctuations to measure. Croxton and Cowden have suggested the following methods.

i. Residual Method
ii. Reference cycle analysis method
iii. Direct method
iv. Harmonic analysis method

Out of these methods only residual method is most commonly used. This method is based on multiplicative model. With the help of this model eliminate trend and after that seasonality is also eliminated. Thus,

$$\frac{T \times S \times C \times I}{T} = \frac{S \times C \times I}{S} = C \times I$$

If there are no irregular fluctuations, the obtained figures represent cyclical variations but if these values have irregular variations than those are to be removed by any of the suitable method as mentioned above. Generally moving average method is used.

14.4 Measurement of Irregular Variations

These variations are accidental, random or simply due to chance factors. These fluctuations, may be caused by such isolated incidents like famines, flood, strikes, wars or sudden changes in demand. By their very nature these movements are very irregular and unpredictable. Thus quantitatively it is almost impossible to separate out the irregular movements and the cyclical movements. So while analyzing time series the trend and seasonal variations are measured separately and the cyclical and irregular variations are left altogether. Thus the irregular component in a time series, represents the residue of variations after trend, seasonal, cyclical movements have been accounted for.

i. If Multiplicative model is used then the original data is divided by T, S. C and we will get irregular variations.

ii. Generally, additive method is used for measurement of random variations.

Thus I = O - T - (S+C)

iii. Moving average method (as mentioned in measurement of seasonal variations) is also used to calculate irregular fluctuations.

Illustration: 9

Using the data given below, calculate seasonal variations and irregular fluctuations:

Year	Summer	Monsoon	Autumn	Winter
2002	30	81	62	119
2003	33	104	86	171
2004	42	153	99	221
2005	56	172	129	235
2006	67	201	136	302

Solution

Calculation of Seasonal Variations and Irregular Fluctuations

Year	*Quarters*	*Values*	*Quarterly Total*	*Total Centered*	*Moving Average*	*Short-time Oscillations (iii - iv)*	*Seasonal Variations*	*Irregular Fluctuations (vii-viii)*
2002	Summer	30	-	-	-	-	-	-
	Monsoon	81	-	-	-	-	-	-
			292					
	Autumn	62		587	73	-11	-19	+8
			295					
	Winter	119		613	77	+42	+68	-26
			318					
2003	Summer	33		660	83	-50	-75	+25
			342					
	Monsoon	104		736	92	+12	+25	-13
			394					
	Autumn	86		797	100	-14	-19	+5
			403					
	Winter	171		855	107	+64	+68	-4
			452					
2004	Summer	42		917	115	-73	-75	+2
			465					
	Monsoon	153		980	123	+30	+25	+5
			515					
	Autumn	99		1044	131	-32	-19	-13
			529					
	Winter	221		1077	135	+86	+68	+18
			548					
2005	Summer	56		1126	141	-85	-75	-10
			578					
	Monsoon	172		1170	146	+26	+25	+1
			592					
	Autumn	129		1195	149	-20	-19	-1
			603					
	Winter	235		1235	154	+81	+68	+13
			632					
2006	Summer	67		1271	159	-92	-75	-17
			639					
	Monsoon	201		1345	168	+33	+25	+8
			706					
	Autumn	136		-	-	-	-	-
	Winter	302	-	-	-	-	-	-

Seasonal variations shown in column viii have been calculated as follows:-

Computation of Average Seasonal Variations

Year	Summer	Monsoon	Autumn	Winter
2002	-	-	-11	+42
2003	-50	+12	-14	+64
2004	-73	+30	-32	+86
2005	-85	+26	-20	+81
2006	-92	+33	-	-
Total	-300	+101	-77	+273
Average	-75	+25	-19	+68

14.5 Summary

Seasonal variation involves patterns of changes that repeat over a period of one year or less. Some factors like seasons, climate, customs or festivals causes these variations. There are many techniques available for computing an index of seasonal variations. Simple average method is the simplest method but it assumes that there is no trend component in the time series, which is not a justified assumption. Moving average method is based on additive model which helps in calculating all short term variations i.e. seasonal and irregular fluctuations. Ratio to trend method measures variations for each month or season for which data are available, so there is no loss of data as occurs in the case of moving average. This is a distinct advantage, when the period of time series is very short. Ratio to moving average method is the most widely used method of measuring seasonal variations. A seasonal index may be used for economic forecasting and managerial control.

Cyclical fluctuations are long term movements that represent consistently, recurring rises and declines in an activity. There are four periods in a business cycle, namely - (i) prosperity (ii) decline (iii) depression and (iv) improvement. The study of cyclical variation is extremely useful like, for avoiding periods of booms and depressions as both are bad for an economy. Despite the importance of business cycles, they are the most difficult type of economic fluctuations to measure.

The irregular component represents the residue of fluctuations after trend, cyclical and seasonal movements have been accounted for. They are caused by such isolated factors like floods, earthquakes, strikes etc.

14.6 Key Words

Seasonal Variation: The seasonal variation in a time series is the repositive, recurrent pattern of change which occurs within a year or shorter time period.

General Average: If seasonal averages are divided by number of seasons or no of months, the resultant is general average.

Link Relative: If the value of current season or month is divided by the value of previous season or month, this is called link relative.

Ratio to Moving Average: If the original values are expressed as percentage of the corresponding moving average values, this is ratio to moving average.

Cyclical Variation: Long term movements that represents consistently recurring rises and declines around secular trend levels, which have a duration normally from 2 to 15 years.

Irregular Variations: Random or accidental variations which do not repeat in a definite pattern.

Ratio to Trend: If original data is divided by corresponding trend values and multiplied by 100, this is called ratio to trend.

14.7 Self Assessment Questions

Q. 1 What is meant by seasonal variation of a time series?

Q. 2 Describe the relative significance of different methods to measure seasonal variations.

Q. 3 Distinguish between seasonal variations and cyclical fluctuations.

Q. 4 What do you mean by short term oscillations? Explain the terms regular and irregular fluctuations.

Q. 5 Write short notes on the following -

i. Seasonal Variation
ii. Irregular Variation
iii. Ratio to Trend Method.

Q. 6 Give the procedure of calculating seasonal variations by link relatives method.

Q. 7 Calculate seasonal indices from the following data using the seasonal average method.

Year	I Qtr.	II Qtr.	III Qtr.	IV Qtr.
2000	72	68	80	70
2001	76	70	82	74
2002	74	66	84	80
2003	76	74	84	78
2004	78	74	86	82

Ans. Seasonal Index (98.43, 92.15, 108.9, 100.52)

Q. 8 By using the ratio to moving average method calculate the seasonal index for each seasons.

Year	Summer	Monsoon	Autumn	Winter
2005	200	180	185	95
2006	220	188	173	83
2007	220	176	161	87

Ans. (131.34, 109.87, 106.29, 52.5)

Q. 9 Using the data given below, calculate the seasonal variations through moving average method.

Year	I Qtr.	II Qtr.	III Qtr.	IV Qtr.
2005	75	60	54	59
2006	86	65	63	80
2007	90	72	66	85
2008	100	78	72	93

Ans. I-(16.4), II-((-6.07), III-(-11.3), IV-(0.6)

Q. 10 Find seasonal variation by ratio to trend method from the data given below :-

Year	Summer	Monsoon	Autumn	Winter
2004	30	40	36	34
2005	34	52	50	44
2006	40	58	54	48
2007	54	76	68	62
2008	80	92	86	82

Ans. (92.05, 117.36, 102.12, 88.46)

Q. 11Find out the seasonal variations by chained index method:-

Year	I Qtr.	II Qtr.	III Qtr.	IV Qtr.
2005	10	50	70	80
2006	20	40	30	90
2007	30	60	40	120
2008	40	20	80	40

Ans. (30.44, 62.10, 102.7, 204.70)

Q. 12Calculate seasonal indices from the data given below by link relatives method:

Year	**I Qtr.**	**II Qtr.**	**III Qtr.**	**IV Qtr.**
2005	68	62	61	63
2006	65	58	66	61
2007	68	63	63	67
2008	72	60	62	65
2009	65	55	60	60

Ans. (106.37, 94.09, 98.97, 100.57)

14.8 Reference Books

1. Gupta, S.P., Statistical Methods.
2. Sharma, J.K., Business Statistics.
3. Agarwal, D.R., Business Statistics.
4. Pinnai & Bhagwati, Statistical Methods.
5. Nagar, K.N., Statistical Methods.
6. Yadav, Jain, Mittal, Statistical Methods.

Unit - 15 Interpolation and Extrapolation - I

Structure of Unit:

15.0 Objectives

After completing this unit, you will be able to :

- Define and differentiate Interpolation and Extrapolation.
- Understand significance, assumptions and accuracy of Interpolation and Extrapolation.
- Use Direct Binomial Expansion Method and Newton's Advancing Difference Method.

15.1 Introduction and Meaning

Sometimes, it may happen due to unavoidable circumstances that intermediate value of a series remains unknown. In such a situation it becomes necessary to estimate or calculate missing figure(s) to analyse and interpret the series so that one may draw accurate conclusions. Interpolation method helps to compute missing values lie between two extreme points. In case, a value lies outside the two extreme value we have to adopt extrapolation method. For example, we are given census figure for 1951, 1961, 1971, 1981, 1991, 2001. If population figures for 1968 and 1994 are required then we can compute the figures with the help of interpolation method but in case population figures for 1945 and 2005 are required then extrapolation method will be used to compute the figures. In the words of W.M. Harper, "Interpolation consists in reading a value which lies between two extreme points, extrapolation means reading a value that lies outside the two extreme points". Interpolation provides information regarding past data within the given series while extrapolation provides information regarding past as well as future data out side the given series.

15.2 Significance of Interpolation and Extrapolation

1. **Estimation for intercensal years :** In case data relating to a particular nature are collected at regular interval and data for different intercensal years are required. Interpolation is only the method which is used to compute the data required. Census of population is the best example in this connection.

2. **Helpful in filling up gaps :** When collected data are insufficient and there is a gap in coverage then it will be appropriate to obtain data with the help of interpolation and extrapolation.

3. **Fill up of destroyed or lost data :** It the collected data are lost or destroyed by fire, flood, earthquake and other natural calamities then interpolation method is used.

4. **Future estimation :** Need of future data for political, social and economic purposes can be satisfied by using available present data. Extrapolation method is adopted to estimate future data.

5. **Determination of positional averages :** Interpolation method is used to determine the value of median and mode in continuous series.

6. **Comparative study :** When it becomes difficult to compare data of different institutions or countries, then interpolation and extrapolation methods are used to make them comparable.

Interpolation and extrapolation are not only useful in statistical research investigations but also useful in diversified field of life. On the basis of best judgement of future trends, governments formulate tax policies, industrial policies. Estimation of future data is useful for preparing budgets. The success of modern business depends on accuracy of estimations and forecasting. Changes to be taken place in future demand and production can be estimated and these may help in planning of production and sales. There are so many business decision areas where interpolation and extrapolation are useful. Economists, sociologists, planning experts, financial analysts, politicians etc. also use these methods.

15.3 Assumptions

1. **No sudden jump :** There is a normal regularity in the data. There was no extra ordinary event e.g. war, flood, earthquake etc. during the period of data available. In other words, it can be said that there are no sudden ups and downs in the data.

2. **Uniformity in rate of changes :** It is assumed that the rate of change of figure is uniform in the given period. There is a continuity in increase or decrease.

15.4 Accuracy of Interpolation and Extrapolation

Interpolated and extrapolated values are only most likely values under certain assumptions, therefore, it should not be expected that these values will be accurate as actual values. The assumptions of interpolation and extrapolation may not hold good in practice in many cases, therefore, our estimates will not be accurate. According to Dr. Bowley, "The accuracy of interpolation depends (i) on knowledge of the possible fluctuations of the figures to be obtained by a general inspection of the fluctuations at dates for which they are given; (ii) on knowledge of the course of the events will which the figures are connected". Accuracy of interpolation and extrapolation also depends on selection of a suitable method.

15.5 Methods of Interpolation and Extrapolation

The some of more popular and important algebric methods are :

(1) Direct Binomial Expansion Method
(2) Newton's Method of Advancing Differences
(3) Newton's Method of Divided Differences
(4) Lagrange's Method
(5) Parabolic Curve Method

In this unit we will discuss only Direct Binomial Expansion Method and Newton's Method of Advancing differences. The remaining three methods will be discussed in the next chapter.

(I) **Direct Binomial Expansion Method :** This method is based on Binomial Theorem. It is used in the situations when the following two conditions are fulfilled :

(i) Independent variable (x) advances by equal intervals, for example, 1990, 1995, 2000, 2005, 2010 and so on. If the difference in independent variable is not uniform, this method can not be used. For example, if x is 1990, 1992, 1995, 1999, 2006 and 2007 .....

(ii) The value of independent variable (x) for which we are interpolating the y, must be one of the class limits of x series.

For example, Look at the following data :

x (year)	:	1951	1961	1971	1981
y (population in crore) :		34	42	?	80

We can interpolate the population (y) of x = 1971, but not corresponding to x = 1965 or 1975.

The same situation is true for extrapolation i.e., we can extrapolate the population for x = 1991 but not for x = 1985.

Procedure :

(1) Firstly, subscript for the independent (x) variable serially $x_0, x_1, x_2, x_3 \ldots\ldots\ldots x_n$ and same for dependent variable (y) = $y_0, y_1, y_2, y_3 \ldots\ldots\ldots y_n$.

(2) Then, the first subscript of dependent variable (y) will be the number for which we have to know the binomial expansion.

If $(y-1)^6 = 0$, Then first y will be y_6 and will be reduced by 1 till it reaches y_0

$y_6 \quad y_5 \quad y_4 \quad y_3 \quad y_2 \quad y_1 \quad y_0$

(3) Both the signs plus and minus are to be placed alternatively. Starting will be plus :

$+y_6 \quad -y_5 \quad +y_4 \quad -y_3 \quad +y_2 \quad -y_1 \quad +y_1$

(4) Then numerical coefficients for y will be given by applying following formula :

$$\frac{\text{Coefficien t of previous y} \times \text{Subscript of previous y}}{\text{Serial No. (relative position) of previous y}}$$

The numerical coefficient can also be found by 'Pascal's Triangle', given below :

Pascal's Triangle

(Power)	**(Numerical Coefficient)**											**2^n**
n												
1						1		1				2
2					1		2		1			4
3				1		3		3		1		8
4			1		4		6		4		1	16
5		1		5		10		10		5		1 32
6	1		6		15		20		15		6	1 64
7		1		7		21		35		35		21 7 1 128
8	1		8		28		56		70		56	28 8 1 256
9		1		9		36		84		126		126 84 36 9 1 512
10	1		10		45		120		210		252	210 120 45 10 1 1024

When we expand the binomial expansion $(y-1)^n$ and equate it to zero.

The expanded form will be as follows :

$$(y-1)^n = y^n - ny^{n-1} + \frac{n(n-1)}{2!}y^{n-2} - \frac{n(n-1)(n-2)}{3!}y^{n-3} + \ldots\ldots\ldots = 0$$

(where n is the number of known value of y).

For example,

If known values of y is (n) 4 then,

$\Delta^4_0 = (y-1)^4 = 0$

$$= y^4 - 4y^{4-1} + \frac{4(4-1)}{1\times 2}y^{4-2} - \frac{4(4-1)(4-2)}{1\times 2\times 3}y^{4-3} + \frac{4(4-1)(4-2)(4-3)}{1\times 2\times 3\times 4}y^{4-4} = 0$$

$$= y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$$

Some Binomial expansion are as follows :

No. of known values	Basic formula	Binomial expansion
2	$(y-1)^2 = 0$	$y_2 - 2y_1 + y_0 = 0$
3	$(y-1)^3 = 0$	$y_3 - 3y_2 + 3y_1 - y_0 = 0$
4	$(y-1)^4 = 0$	$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$
5	$(y-1)^5 = 0$	$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$

Illustration 1 :

Estimate the production for the year 1975 with the help of following table :

x	1965	1970	1975	1980	1985	1990	1995
y (Production In Tonnes)	1331	1728	2197	2744	3375	4096	4913

Solution :

Since the known values are six, the sixth leading differences will be zero, i.e. $(y-1)^6 = 0$ or $\Delta_0^6 = 0$

$$\Delta_0^6 = y_6 - 6y_5 + 15y_4 - 20y_3 + 15y_2 - 6y_1 + y_0 = 0$$

Year		Production	
1965	x_0	1331	y_0
1970	x_1	1728	y_1
1975	x_2	2197	y_2
1980	x_3	2744	y_3
1985	x_4	3375	y_4
1990	x_5	4096	y_5
1995	x_6	4913	y_6

Substituting the value

$$4913 - (6 \times 4096) + (15 \times 3375) - (20x2744 + (15yl) - (6 \times 1728) + 1331 = 0$$

or $4913 - 24576 + 50625 - 54880y_3 + 15yl - 10368 + 1331 = 0$

or $-15y_3 - 32955 = 0$

or $15y_3 = 32955$

or $y_2 = 2197$

Hence the estimated production for the year 1980 is 2744 tonnes.

Illustration 2 :

The age of mother and the average number of children born per mother are given in the following table. Using appropriate formula, find the average number of children born per mother aged 30-34 years.

Age of mother (In years)	15-19	20-24	25-29	30-34	35-39	40-44
Average number of children born	0.7	2.1	3.1	?	5.7	5.8

(B.Com., Madras Univ., 1998)

Solution :

We shall use Binomial Expansion Method since the known figures are five, fifth leading differences will be zero.

$$\Delta^5_0 = y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

Age of mother (In years)	15-19 x_0	20-24 x_1	25-29 x_2	30-34 x_3	35-39 x_4	40-44 x_5
Average number of children born	0.7 y_0	2.1 y_1	3.1 y_2	? y_3	5.7 y_4	5.8 y_5

Substituting the given values

$\Delta^5_0 = (5.8) - (5 \times 5.7) + (10y_3) - (10 \times 3.1) + (5 \times 2.1) - 0.7 = 0$

$= 5.8 - 28.5 + 10Y_3 - 31 + 10.5 - 0.7 = 0$

$10y_3 = 28.5 - 5.8 + 31 - 10.5 + 0.7$

$10y_3 = 43.9$

$y_3 = 4.39$

Hence, the average number of children born per mother aged 30-34 is 4.39

Two or more missing values.

Binomial expansion can be easily used in the situation where two or more values are missing. In a series, when two values are missing, we are having two unknown quantities in the equation obtained by binomial expansion. In this situation we need two binomial expansion equations.

For example : In a case we are given 5 known values and have to interpolate 2 unknown values, then following two equations will be as follows :

$\Delta^5_0 = y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$

$\Delta^5_1 = y_6 - 5y_5 + 10y_4 - 10y_3 + 5y_2 - y_1 = 0$

with the help of two equations, we will interpolate the missing values.

Illustration 3 :

Working class cost of living indices of a city are given below. Find the index number for the year 2006.

Year	2004	2005	2006	2007	2008
Index No.	200	214	-	314	424

Solution :

Here we can apply binomial expansion method to estimate the index number for the year 2006.

Year	2004	2005	2006	2007	2008
Index No.	200 y_0	214 y_1	– y_2	314 y_3	424 y_4

Since the known values are four, the fourth leading differences will be zero.

$\Delta_0^4 = y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$

Putting the given values.

$424 - 4(314) + 6y_2 - 4(214) + 200 = 0$

$424 - 1256 + 6y_2 - 856 + 200 = 0$

$6y_2 = 1256 + 856 - 424 - 200$

$6y_2 = 1488$

$y_2 = 248$

Hence, the index number for the year 2006 is 248.

Illustration 4 :

Estimate the production of sugar for the year 2005 from the following data :

Year	2003	2004	2005	2006	2007	2008
Production (m. tonnes)	640	600	–	560	556	500

Solution :

As five values are known, the fifth leading difference will be zero. i.e. $\Delta_0^5 = 0$.

Year	2003	2004	2005	2006	2007	2008
Production (m. tonnes)	640 y_0	600 y_1	– y_2	560 y_3	556 y_4	500 y_5

$$D_0^5 = y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

Substituting the values.

$$500 - 5(556) + 10(560) - 10y_2 + 5(600) - 640 = 0$$
$$500 - 2780 + 5600 - 10y_2 + 3000 - 640 = 0$$
$$-10y_2 = 2780 + 640 - 500 - 5600 - 3000$$
$$-10y_2 = -5680$$
$$10y_2 = 5680$$
$$y_2 = 568$$

Hence the production for 2005 is 568 m.tonnes.

Illustration 5 :

Interpolate the missing values in the following table —

x	20	21	22	23	24	25	26
y	135	?	111	100	?	82	74

(B.A. Raj. Univ. 1993)

Solution :

x	20 x_0	21 x_1	22 x_2	23 x_3	24 x_4	25 x_5	26 x_6
y	135 y_0	? y_1	111 y_2	100 y_3	? y_4	82 y_5	74 y_6

Since the known values are five, the fifth leading difference will be zero. Two values are missing. So, Two Binomial expansion will be as follows.

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0 \qquad (1)$$
$$y_6 - 5y_5 + 10y_4 - 10y_3 + 5y_2 - y_1 = 0 \qquad (2)$$

Substituting the values in two equations.

$$82 - 5y_4 + (10 \times 100) - (10 \times 111) + 5y_1 - 135 = 0$$
$$74 - (5 \times 82) + 10y_4 - (10 \times 100) + (5 \times 111) - y_1 = 0$$
$$5y_1 - 5y_4 = 163 \qquad (3)$$
$$-y_1 + 10y_4 = 781 \qquad (4)$$

Multiply the (4) equation by 5 and then add in (3) equation

$$5y_1 - 5y_4 = 163$$
$$-5y_1 + 50y_4 = 3905$$
$$45y_4 = 4068$$
$$y_4 = 90.4$$

Putting the value of y_4 in (4) equation

$-y_1 + 904 = 781$

$-y_1 = -123$

$y_1 = 123$

Hence, the value for x = 21 and 24 is 123 and 90.4

Illustration 6 :

From the following data, estimate the sales for the years 2004 and 2006.

Year	Sales (in Tonnes)
2001	100
2002	110
2003	130
2004	–
2005	175
2006	–
2007	215

Solution :

Since there are five known values, the fifth leading difference will be zero. i.e. $D_0^5 = 0$.

There are two unknown values of sales. Therefore, two equations will be used to determine them. The equations are :

$\Delta_0^5 = y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$

$\Delta_1^5 = y_6 - 5y_5 + 10y_4 - 10y_3 + 5y_2 - y_1 = 0$

Year	2001	2002	2003	2004	2005	2006	2007
Sales	100 y_0	110 y_1	130 y_2	– y_3	175 y_4	– y_5	215 y_6

Putting the values,

$y_5 - 5(175) + 10y_3 - 10(130) + 5(110) - 100 = 0$ .....(i)

$215 - 5y_5 + 10(175) - 10y_3 + 5(130) - 110 = 0$ .....(ii)

or $y_5 - 875 + 10y_3 - 1300 + 550 - 100 = 0$

$215 - 5y_5 + 1750 - 10y_3 + 650 - 110 = 0$

or $y_5 + 10y_3 = 1725$

$-5y_5 - 10y_3 = -2505$

$4y_5 = 780$

$y_5 = 195$

Putting the value of y_5 in above equation,

$y_5 + 10y_3 = 1725$

$195 + 10y_3 = 1725$

$10y_3 = 1530$

$y_3 = 153$

Hence the sales for the year 2004 and 2006 are 153 and 195 tonnes respectively.

Illustration 7 :

Working class cost of living indices of a certain place for some years are given below. Extrapolate the index number for 1999.

Year	1994	1995	1996	1997	1998
Index No.	150	235	365	525	780

Solution :

We can extrapolate the living index for 1999 by Binomial Expansion method. The known values are five, the fifth leading difference will be zero,

$(y-1)^5 = 0$

or

$\Delta^5_0 = 0$

$\Delta^5_0 = y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$

x (Year)	1994 x_0	1995 x_1	1996 x_2	1997 x_3	1998 x_4	1999 x_5
y (Index No.)	150 y_0	235 y_1	365 y_2	525 y_3	780 y_4	? y_5

Putting the values in equation,

$(y-1)^5 = y_5 - (5 \times 780) + (10 \times 525) - (10 \times 365) + (5 \times 235) - 150 = 0$

$= y_5 - 3900 + 5250 - 3650 + 1175 - 150 = 0$

$- y_5 = 3900 - 5250 + 3650 - 1175 + 150$

$- y_5 = - 1275$

$y_5 = 1275$

Hence the estimated index no. for 1999 is 1275

In the situation of an extra ordinary value.

The situation where in a value in given values is extra ordinary value will be replaced by new interpolated value. Further procedure is same as explained earlier.

Illustration 8 :

The following data give the profit of a firm. Find the profit for the year 2006

Year	2001	2002	2003	2004	2005
Profit (in crore Rs.)	7	9	36	14	16

Solution :

It is clear by observation that profit for the year 2003 is extra ordinary. So, first of all we will interpolate the profit for 2003.

Year (x)	2001 x_0	2002 x_1	2003 x_2	2004 x_3	2005 x_4
Profit (y)	7 y_0	9 y_1	? y_2	14 y_3	16 y_4

Since the known values are four. so the fourth leading difference will be zero.

$\Delta^4_0 = 0$

$\Delta^4_0 = y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$

Putting the known values in the formula :

$16 - (4 \times 14) + 6y_2 - (4 \times 9) + 7 = 0$

or $16 - 56 + 6y_2 - 36 + 7 = 0$

or $6y_2 - 69 = 0$

or $y_2 = 11.5$

So the interpolated value for the year 2003 is Rs. 11.5 crore with the help of this interpolated value, we will extrapolate the profit for the year 2006.

Extrapolation for the year 2006

Year (x)	2001 x_0	2002 x_1	2003 x_2	2004 x_3	2005 x_4	2006 x_5
Profit (y) (In Rs. Crore)	7 y_0	9 y_1	11.5 y_2	14 y_3	16 y_4	? y_5

For the series, known value are five, so the fifth leading difference will be zero.

$\Delta^5_0 = 0$

$\Delta^5_0 = y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$

Putting the values in the formula

$y_5 - (5\times16) + (10\times14) - (10\times11.5) + (5\times9) - 7 = 0$

or $y_5 - 80 + 140 - 115 + 45 - 7 = 0$

$y_5 - 17 = 0$

$y_5 = 17$

So the profit for the year 2006 is Rs. 17 crore

Activity A

1. Sohan has to interpolate production for a particular year. he is given production (in tonnes) for 1991, 1992, 1993, 1994, 1996, 1997. He has to interpolate the production for 1995. Help him in expanding the binomial equation when known values are six.

   _____

   _____

   _____

2. Ram has given a question by his teacher. He has to write two Binomial equations when 7 known values are given and 2 values are missing. Help Ram in writing the equations.

   _____

   _____

   _____

(II) Newton's Advancing Differences Method : Newton's Advancing Difference Method is also based on Binomial Theorem. This method can be used only when the values of the argument (independent variable) X are equidistant. But in this method it is not necessary that the value of X for which y is to be interpolated is one of the class limits of x series.

For example :

x	y
1961	50 y_0
1971	60 y_1
1981	70 y_2
1991	80 y_3
2001	90 y_4

By the use of this method we can interpolate the value of y for x = 1965 or 1974 etc. Similarly, we can extrapolate the value for x = 2005, 2009 etc.

Procedure

In Newton's Advancing Difference Method, procedure is as follows :

(i) Subscript for the independent variable x serially i.e. x_0, x_1, x_2, x_3, x_4 ..... .

(ii) Subscript for the dependent variable y serially i.e. y_0, y_1, y_2, y_3, y_4 ..... .

(iii) Prepare table of difference to find leading Differences.

Format of Table of Difference is given below :

Table showing Finite or Advancing Difference

X (Independent Variable)	Y (Dependent Variable)	Differences			
		First Difference Δ^1	Second Difference Δ^2	Third Difference Δ^3	Fourth Difference Δ^4
x_0	y_0	$y_1 - y_0$ Δ_0^1	$\Delta_1^1 - \Delta_0^1$ Δ_0^2	$\Delta_1^2 - \Delta_0^2$ Δ_0^3	$\Delta_1^3 - \Delta_0^3$ Δ_0^4
x_1	y_1	$y_2 - y_1$ Δ_1^1	$\Delta_2^1 - \Delta_1^1$ Δ_1^2	$\Delta_2^2 - \Delta_1^2$ Δ_1^3	
x_2	y_2	$y_3 - y_2$ Δ_2^1	$\Delta_3^1 - \Delta_2^1$ Δ_2^2		
x_3	y_3	$y_4 - y_3$ Δ_3^1			
x_4	y_4				

The differences are indicated by sign Δ are to be calculated. The first difference in each column in called as leading differences. Thus, the first difference is indicated by Δ^1_0, second difference by Δ^2_0, third difference by Δ^3_0 fourth difference by Δ^4_0.

It is clear by observing this table that if all leading differences are known, we can calculate remaining differences and values for y.

Formulae :

$$y_1 = y_0 + \Delta_0^1$$

$$y_2 = y_1 + \Delta_1^1 = y_0 + \Delta_0^1 + \Delta_0^2 + \Delta_0^1$$

$$y_3 = y_2 + \Delta_2^1 = y_0 + 3\Delta_0^1 + 3\Delta_0^2 + \Delta_0^3$$

(Special attention should be paid for algebric signs (+ and –) while calculating difference table

(iv) Calculate difference of independent variable x : After calculating leading differences, the value of x should be calculated. The value of x shall be obtained as follows.

$$x = \frac{\text{The value to be interpolated} - \text{The value at origin}}{\text{Difference between the two adjoining Items}}$$

or $$x = \frac{x_x - x_0}{x_1 - x_0}$$

(v) Lastly, use Newton's formula by putting the values in formula, calculate the interpalated value. Newton's advancing formula is as follows :

$$y_x = y_0 + x\,\Delta_0^1 + \frac{x(x-1)}{1\times 2}\,\Delta_0^2 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\,\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\,\Delta_0^4 + \ldots\ldots\ldots\ldots$$

Applicability :

Newton's advancing difference method should be used when the value to be interpolated is in the beginning of the table. The formula used in the method considers only leading differences which are always in the beginning.

Illustration 9 :

The following are the annual premiums charged by the LIC of India for a policy of Rs. 1000. calculate the premium payable at the age of 26.

Age in years	20	25	30	35	40
Premium (Rs.)	23	26	30	35	42

(MBA, HPU)

Solution :

In this situation, we will apply Newton's advancing difference method,

Advancing Difference Table

Age in Years (x)		Premiums (y)		First Difference Δ^1		Second Difference Δ^2		Third Difference Δ^3		Fourth Difference Δ^4	
20	x_0	23	y_0								
				+3	Δ_0^1						
25	x_1	26	y_1			+1	Δ_0^2				
				+4	Δ_1^1			0	Δ_0^3		
30	x_2	30	y_2			+1	Δ_1^2			1	Δ_0^4
				+5	Δ_2^1			1	Δ_1^3		
35	x_3	35	y_3			+2	Δ_2^2				
				+7	Δ_3^1						
40	x_4	42	y_4								

$$y_x = y_0 + x\,\Delta_0^1 + \frac{x(x-1)}{1\times 2}\,\Delta_0^2 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\,\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\,\Delta_0^4$$

$$x = \frac{26-20}{5} = 1.2$$

Putting the values in the formula

$$y_{26} = 23 + (1.2\times 3) + \frac{1.2(1.2-1)}{1\times 2}\times 1 + \frac{1.2(1.2-1)(1.2-2)}{1\times 2\times 3}\times 0 + \frac{1.2(1.2-1)(1.2-2)(1.2-3)}{1\times 2\times 3\times 4}\times 1$$

$$y_{26} = 26.734$$

Thus, the premium payable on Rs. 1000 at the age of 26 is Rs. 26.73.

Illustration 10 :

The following table gives the population of a town for different census years, Estimate the population of the town for the year. 1985.

Year	1961	1971	1981	1991	2001
Population (In Lakh)	70	90	360	140	160

Solution :

By observing the given data, we found that there is some mistake in the population of year 1981 (because it is extra ordinary high). So, first of all, using Binomial method, we will interpolate the population of 1981. Then, including this interpolated population. We will interpolate the population for the year 1985 by Newton's advancing difference method.

x		**y**	
1961	x_0	70	y_0
1971	x_1	90	y_1
1981	x_2	?	y_2
1991	x_3	140	y_3
2001	x_4	160	y_4

Since the known values are four, so the fourth leading difference will be zero.

$$\Delta^4{}_0 = y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$$

Putting the values in the formula

$$160 - 4\times 140 + 6y_2 - 4\times 90 + 70 = 0$$

$$6y_2 + 230 - 920 = 0$$

$$\therefore \quad y_2 = \frac{690}{6} = 115$$

Now, replacing 360 by 115. Using Newton'a advancing difference method, we will interpolate for the year 1985.

Advancing Difference Table

Years (x)	Population (y)	First Difference Δ^1	Second Difference Δ^2	Third Difference Δ^3	Fourth Difference Δ^4
1961 x_0	70 y_0				
		+20 Δ_0^1			
1971 x_1	90 y_1		+5 Δ_0^2		
		+25 Δ_1^1		–5 Δ_0^3	
1981 x_2	115 y_2		0 Δ_1^2		0 Δ_0^4
		+25 Δ_2^1		–5 Δ_1^3	
1991 x_3	140 y_3		–5 Δ_2^2		
		+20 Δ_3^1			
2001 x_4	160 y_4				

$$x = \frac{x_x - x_0}{x_1 - x_0} = \frac{1985-1961}{1971-1961} = \frac{24}{10} = 2.4$$

$$y_x = y_0 + x\,\Delta_0^1 + \frac{x(x-1)}{1\times 2}\,\Delta_0^2 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\,\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\,\Delta_0^4$$

By putting the values

$$y_x = 70 + (2.4 \times 20) + \left(\frac{2.4\times 1.4}{2}\times 5\right) + \left(\frac{2.4\times 1.4\times 0.4}{1\times 2\times 3}\times -5\right) + \left(\frac{2.4\times 1.4\times 0.4\times -0.6}{1\times 2\times 3\times 4}\times 0\right)$$

$y_x = 70 + 48 + 8.4 - 1.12 + 0$

$y_x = 125.28$ or 125

So, the population for the year 1985 is 125 Lakhs.

Illustration 11 :

The following table relates to the daily income of workers of a company. Interpolate the number of workers income below 45.

Income (Rs.)	**No. of workers**
30–40	31
40–50	42
50–60	51
60–70	35
70–80	31

Solution :

Advancing Difference Table

Income less than (x)	No. of workers (y)	Differences			
		First Difference Δ^1	Second Difference Δ^2	Third Difference Δ^3	Fourth Difference Δ^4
40 x_0	31 y_0				
		+42 Δ_0^1			
50 x_1	73 y_1		+9 Δ_0^2		
		+51 Δ_1^1		–25 Δ_0^3	
60 x_2	124 y_2		–16 Δ_1^2		+37 Δ_0^4
		+35 Δ_2^1		+12 Δ_1^3	
70 x_3	159 y_3		–4 Δ_2^2		
		+31 Δ_3^1			
80 x_4	190 y_4				

$$x = \frac{x_x - x_0}{x_1 - x_0} = \frac{45-40}{50-40} = 0.5$$

$$y_x = y_0 + x\,\Delta_0^{1} + \frac{x(x-1)}{1\times 2}\,\Delta_0^{2} + \frac{x(x-1)(x-2)}{1\times 2\times 3}\,\Delta_0^{3} + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\,\Delta_0^{4}$$

By putting the values in the formula

$$y_x = 31 + (0.5 \times 42) + \left(\frac{0.5\times -0.5}{2}\times 9\right) + \left(\frac{0.5\times -0.5\times -1.5}{1\times 2\times 3}\times -25\right) + \left(\frac{0.5\times -0.5\times -1.5\times -2.5}{1\times 2\times 3\times 4}\times 37\right)$$

$y_x = 31 + 21 - 1.125 - 1.5625 - 1.4453$

$y_x = 47.8672$ or 48

So the No. of workers below income of Rs. 45 is 48

Illustration 12 :

From the following data estimate the number of students getting marks in statistics more than 75.

Marks Below	No. of Students
50	50
60	150
70	300
80	500
90	700
100	800

Solution :

Applying Newton's Advancing difference Method to ascertain the number of students getting more than75 marks in statistics.

Advancing Difference Table

Marks (x)		No. of student (y)		Differences									
				First Difference Δ^1		Second Difference Δ^2		Third Difference Δ^3		Four Difference Δ^4		Fifth Difference Δ^5	
Below 50	x_0	50	y_0										
				+100	Δ_0^1								
Below 60	x_1	150	y_1			+50	Δ_0^2						
				+150	Δ_1^1			0	Δ_0^3				
Below 70	x_2	300	y_2			+50	Δ_1^2			–50	Δ_0^4		
				+200	Δ_2^1			–50	Δ_1^3			0	Δ_0^5
Below 80	x_3	500	y_3			0	Δ_2^2			–50	Δ_1^4		
				+200	Δ_3^1			–100	Δ_2^3				
Below 90	x_4	700	y_4			–100	Δ_3^2						
				+100	Δ_4^1								
Below 100	x_5	800	y_5										

$$y_x = y_0 + x\,\Delta_0^{1} + \frac{x(x-1)}{1\times 2}\,\Delta_0^{2} + \frac{x(x-1)(x-2)}{1\times 2\times 3}\,\Delta_0^{3} + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\,\Delta_0^{4} + \frac{x(x-1)(x-2)(x-3)(x-4)}{1\times 2\times 3\times 4\times 5}\,\Delta_0^{5}$$

and $\quad x = \frac{75-50}{10} = 2.5$

By putting the values in the formula

$$y_{75} = 50 + (2.5 \times 100) + \frac{2.5(2.5-1)}{1\times 2} \times 50 + \frac{2.5(2.5-1)(2.5-2)}{1\times 2\times 3} \times 0 + \frac{2.5(2.5-1)(2.5-2)(2.5-3)}{1\times 2\times 3\times 4} \times -50 + \frac{2.5(2.5-1)(2.5-2)(2.5-3)(2.5-4)}{1\times 2\times 3\times 4\times 5} \times 0$$

$y_{75} = 50 + 250 + 93.75 + 0 + 1.95 + 0$

$y_{75} = 395.7$ or 396

Thus the number of students getting marks upto 75 is 396. The total number of students is 800. Hence, the number of students getting more than 75 is (800 – 396) = 404.

Activity B

1. Monu, is a student of BBA. He has to prepare the table of differences under Newton's advancing differences upto third leading difference (Δ^3_0). Help him in prepairing this table

   _____

   _____

   _____

2. Anu is a student. She has given a home work by her teacher. Assist her in writing down Newton's advancing difference formula upto Δ^4_0 for interpolating Y x.

   _____

   _____

   _____

Illustration 13 :

From the following table, estimate the number of students getting second division in the examination.

Marks (Out of 100)	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
No. of students	10	52	170	108	60

(48% and above but less than 60% marks make second division)

Solution :

Here we would be know the number of students obtaining less then 48% marks. We would use Newton's method here.

Advancing Difference Table

Marks (less than) (x)		No. of students (y)		Differences							
				First Difference Δ^1		Second Difference Δ^2		Third Difference Δ^3		Fourth Difference Δ^4	
20	x_0	10	y_0								
				+52	Δ_0^1						
40	x_1	62	y_1			+118	Δ_0^2				
				+170	Δ_1^1			–180	Δ_0^3		
60	x_2	232	y_2			– 62	Δ_1^2			+194	Δ_0^4
				+108	Δ_2^1			+14	Δ_1^3		
80	x_3	340	y_3			–48	Δ_2^2				
				+60	Δ_3^1						
100	x_4	400	y_4								

$$y_x = y_0 + x\Delta_0^1 + \frac{x(x-1)}{1\times 2}\Delta_0^2 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\Delta_0^4$$

$$x = \frac{48-20}{40-20} = 1.4$$

$$y_{48} = 10 + 1.4(52) + \frac{1.4(1.4-1)}{1\times 2}(118) + \frac{1.4(1.4-1)(1.4-2)}{1\times 2\times 3}(-180) + \frac{1.4(1.4-1)(1.4-2)(1.4-3)}{1\times 2\times 3\times 4}(194)$$

$y_{48} = 10 + 72.8 + 33.04 + 10.08 + 4.34$

$y_{48} = 130.36$ or 130

Number of students getting less than 60% marks = 232

Number of students getting less than 48% marks = 130

So, the number of students getting second division (232 – 130) = 102

Illustration 14 :

From the following data, interpolate the number of persons living between the age of 35 and 42.

Age (In year)	20	30	40	50
No. of persons living	1026	878	692	486

Solution :

We would use here Newton's Method

Advancing Difference Table

Marks (less than) (x)		No. of students (y)		Differences					
				First Difference Δ^1		Second Difference Δ^2		Third Difference Δ^3	
20	x_0	1026	y_0						
30	x_1	878	y_1	–148	Δ_0^1				
						–38	Δ_0^2		
40	x_2	692	y_2	–186	Δ_1^1			+18	Δ_0^3
						–20	Δ_1^2		
50	x_3	486	y_3	–206	Δ_2^1				

$$y_x = y_0 + x\Delta_0^1 + \frac{x(x-1)}{1\times 2}\Delta_0^2 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\Delta_0^3$$

Number of persons living at 35,

$$x = \frac{35-20}{30-20} = 1.5$$

Putting the values,

$$y_{35} = 1026 + 1.5(-148) + \frac{1.5(1.5-1)}{1\times 2}(-38) + \frac{1.5(1.5-1)(1.5-2)}{1\times 2\times 3}(18)$$

$y_{35} = 1026 - 222 - 14.250 - 1.125$

$y_{35} = 788.625$ or 789

Number of persons living at 42.

$$x = \frac{42-20}{30-20} = 2.2$$

$$y_{42} = 1026 + 2.2(-148) + \frac{2.2(2.2-1)}{1\times 2}(-38) + \frac{2.2(2.2-1)(2.2-2)}{1\times 2\times 3}(18)$$

$y_{42} = 1026 - 325.6 - 50.16 - 1.58$

$y_{42} = 651.82$ or 652

Number of persons living at age of 42 = 652

Number of persons living at age of 35 = 789

Number of persons living between 35 and 42 is (789 – 652) = 137.

Illustration 15 :

Using Newton's method of interpolation, estimate from the following data, the number of workers earning between Rs. 60 and Rs. 70 per day in a concern.

Earning (in Rs.)	No. of workers
Below 40	500
40 – 60	240
60 – 80	200
80 – 100	140
100 – 120	100

Solution :

Advancing Difference Table

Earning (in Rs.) less than (x)	No. of workers (y)	Differences			
		First Difference Δ^1	Second Difference Δ^2	Third Difference Δ^3	Fourth Difference Δ^4
40 x_0	500 y_0				
		+240 Δ_0^1			
60 x_1	740 y_1		–40 Δ_0^2		
		+200 Δ_1^1		–20 Δ_0^3	
80 x_2	940 y_2		–60 Δ_1^2		+40 Δ_0^4
		+140 Δ_2^1		+20 Δ_1^3	
100 x_3	1080 y_3		–40 Δ_2^2		
		+100 Δ_3^1			
120 x_4	1180 y_4				

$$y_x = y_0 + x\Delta_0^1 + \frac{x(x-1)}{1\times 2}\Delta_0^2 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\Delta_0^4$$

$$x = \frac{70-40}{60-20} = 1.5$$

$$y_{70} = 500 + 1.5(240) + \frac{1.5(1.5-1)}{1\times 2}(-40) + \frac{1.5(1.5-1)(1.5-2)}{1\times 2\times 3}(-20) + \frac{1.5(1.5-1)(1.5-2)(1.5-3)}{1\times 2\times 3\times 4}(40)$$

$y_{70} = 500 + 360 - 15 + 1.25 + 0.936$

$y_{70} = 846.906$ or 846.9 or 847

Number of workers getting less than Rs. 70 = 847.

Number of workers getting less than Rs. 60 = 740.

So, the number of workers getting between Rs. 60 and Rs. 70 is (847 – 740) = 107.

15.6 Summary

Interpolation and extrapolation methods are used to compute value lie between or outside the two extreme points. If data are collected on only regular basis and data for different intercensal years to be computed then interpolation method is used to calculate unknown figure. Gap in coverage data, lost or destroyed due to unavoidable reasons, future estimation, determination of positional averages and need of comparative study are some important factors which allures to compute interpolated and extrapolated values. There are two assumptions of interpolation and extrapolation. The first assumption underlines that there should not be any extra ordinary event during the given period. It may lead to ups and downs in the figures. The second assumption is based on uniformity in rate of changes of figures. Interpolated and extrapolated values are most likely values therefore, these will not be accurate. There are five more popular algebric methods. Direct Binomial Expansion method is used when independent variable (x) advances by equal intervals, the value of independent variable (x) for which we are interpolating the y, must be one of the class limits of x series.

Newton's advancing difference method can be used only when the values of the argument are equidistant. In this method it is not necessary that the values of x for which y is to be interpolated is one of the class limits of x series. The formula used in the method considers only leading differences which are always in the beginning.

15.7 Key Words

Interpolation : The method to compute the value that lies between the two extreme points.

Extrapolation : The method to compute the value that lies outside the two extreme points.

Direct Binomial Expansion Method : The method is used when independent variable advances by

equal intervals and value of independent variable for which we are interpolating the y, must be one of the class limits of x series.

Newton's advancing difference method : The method is used only when the values of the argument are equidistant and it is not necessary that the value of x for which y is to be interpolated is one of the class limits of x series.

15.8 Self Assessment Questions

1. What do you understand by the term Interpolation and extrapolation? Discuss briefly their necessity and usefulness.
2. Differentiate between Interpolation and Extrapolation. what are the assumptions on which methods of interpolation are based ?
3. What are the two essential conditions that must be satisfied for the application of the Binomial Expansion method? Present the coefficient of the Binomial Expansion $(y-1)^{10}$ in Pascal's Triangle form.
4. Find the missing value from the following table :

x	1	2	3	4	5
y	17	–	23	31	47

(B.Com., III Rajasthan Univ., 1999)

[Ans. $y_x = 19.5$]

5. Interpolate the missing figure of the following table with the help of a suitable formula :

Year	1988	1989	1990	1991	1992
Profit (Rs. in Lakhs)	7	10	?	19	27

(B.Com. Allahabad 1993)

[Ans. 13.7 or Rs 14 Lakhs]

6. Interpolate the missing figures from the following :

x	0	5	10	15	20	25
y	7	10	?	18	?	32

(B.Com., Allahabad 1994)

[Ans. 13.5 and 24]

7. Estimate the production for the years 1976 and 1986 from the following data :

Year	1971	1976	1981	1986	1991	1996
Production	180	–	250	–	320	400

(B.Com., Kurukshetra, 1998)

[Ans. 224 and 276]

8. From the following data, find the missing figures :

Year	1989	1990	1991	1992	1993	1994	1995	1996
Output (000 tonnes)	12	15	?	24	29	?	40	46

[Ans. 19.25 and 34.30]

9. From the following table find the value of y when x = 8.

x	1	3	5	7	9
y	20	30	42	58	72

[Ans. 66]

10. Given below are the figures of population of a district for different years. Find the population for 1975 :

Year	1951	1961	1971	1981	1991
Population (lakhs)	7	9	6	14	16

[By Binomial method population of 1971 is 11.5, then for 1975 = 12.53 lakhs]

11. Estimate the expectation of life at the age of 16 years and 22 years from the following data

Age in Years	10	15	20	25	30	35
Expection of life (in year)	35.4	32.2	29.1	26.0	23.1	20.4

(M.Com., Kanpur, 1997)

[Ans. $y_{16} = 31.58$, $y_{22} = 27.85$]

12. From, the following data find the number of student securing less than 45 marks :

Marks	30–40	30–50	30–60	30–70	30–80
No. of student	31	73	124	159	190

(M.A., Ajmer, 1991)

[Ans. 48]

15.9 Reference Books

1. Gupta, S.P. Statistical Methods.
2. Saha, S. Business Statistics.

Unit - 16 Interpolation and Extrapolation - II

Structure of Unit:

16.0 Objectives

After completing this unit, you will be able to :

- Explain remaining methods of Interpolation and Extrapolation.
- Evaluate the difference in various methods.
- Assess the importance and usefulness of each method.
- Calculate missing values or project future value.

16.1 Introduction

As it has been explained earlier in the previous chapter, that interpolation and extrapolation techniques are applied to estimate the missing values or to project future values. Direct Binomial Expansion method is used where the X-variable advances by equal intervals and the value of x for which y is to be interpolated should be one of the class limits of x series, Newton's Advancing Differences method is applicable where the values of x are equidistant but it is not necessary that the value of x for which y is to be interpolated is one of the class limits of x series. There may be some different situations where these both techniques can not be used. The other remaining techniques of interpolation and extrapolation are being explained in this chapter.

16.2 Newton's Divided Difference Method

Divided difference method was also given by Newton. This method is applicable when the independent variable x increases by unequal intervals.

Procedure

1. Subscript the independent variable x serially i.e. x_0, x_1, x_2, x_3,------
2. Also subscript the dependent variable y for example y_0, y_1, y_2, y_3, ------
3. Construct a table of divided differences, format of preparing table is given below :

Table of Divided Differences

x	y	Divided Differences					
		First Difference Δ^1		Second Difference Δ^2		Third Difference Δ^3	
x_0	y_0	$\frac{y_1 - y_0}{x_1 - x_0}$	Δ_0^1	$\frac{\Delta_1^1 - \Delta_0^1}{x_2 - x_0}$	Δ_0^2	$\frac{\Delta_1^2 - \Delta_0^2}{x_3 - x_0}$	Δ_0^3
x_1	y_1	$\frac{y_2 - y_1}{x_2 - x_1}$	Δ_1^1	$\frac{\Delta_2^1 - \Delta_1^1}{x_3 - x_1}$	Δ_1^2		
x_2	y_2	$\frac{y_3 - y_2}{x_3 - x_2}$	Δ_2^1				
x_3	y_3						

The first, second and third leading difference are Δ^1_0, Δ^2_0, Δ^3_0, respectively

4. Apply the given below formula to find the value to be interpolated.

Formula :

$$y_x = y_0 + (x - x_0)\ \Delta_0^1 + (x - x_0)(x - x_1)\ \Delta_0^2 + (x - x_0)(x - x_1)(x - x_2)\ \Delta_0^3$$

Illustration 1 :

Use a suitable method to find the value of y when x = 5 from the following data :

x	2	6	8	9
y	158	110	62	53

Solution :

Since the independent variable is advancing by unequal intervals, so we will have to use Newton's divided difference method.

Interpolating the value of y for x = 5.

By the divided differences method.

x		y		Differences					
				First Difference Δ^1		Second Difference Δ^2		Third Difference Δ^3	
2	x_0	158	y_0	$\frac{110-158}{6-2}$	$-12\Delta_0^1$	$\frac{-24-(-12)}{8-2}$	$-2\Delta_0^2$	$\frac{5-(-2)}{9-2}$	$1\Delta_0^3$
6	x_1	110	y_1	$\frac{62-110}{8-6}$	$-24\Delta_1^1$	$\frac{-9-(-24)}{9-6}$	$5\Delta_1^2$		
8	x_2	62	y_2	$\frac{53-62}{9-8}$	$-9\Delta_2^1$				
9	x_3	53	y_3						

$$y_x = y_0 + (x - x_0)\ \Delta_0^1 + (x - x_0)(x - x_1)\ \Delta_0^2 + (x - x_0)(x - x_1)(x - x_2)\ \Delta_0^3$$

$$y_x = 158 + (5 - 2) \times (-12) + (5 - 2)(5 - 6) \times (-2) + (5 - 2)(5 - 6)(5 - 8) \times 1$$

$$y_x = 158 - 36 + 6 + 9$$

$$y_x = 137$$

16.3 Lagrange's Method

This is a universal method for interpolation and extrapolation. This method has been devised by a famous french mathematician lagrange.

Features :

1. This method can be used for any type of data. Whether our series increases by regular or irregular intervals.

2. This is applicable in any situation, whether the dependent value (y) to be interpolated is in the beginning or at the end.

In practice Lagrange method is used where Binomial expansion method and Newton's advancing difference method can not be used. For example: The population of a city for the year 1991, 1995, 2000, 2001, 2003 are given. We have to extrapolate the population for the year 2006. So, in this situation we use lagrange's method because Binomial expansion method and Newton's advancing difference method can not be used here.

Formula :

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)....(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)....(x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)....(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)....(x_1-x_n)}$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)....(x-x_n)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)....(x_2-x_n)} + y_n \frac{(x-x0)(x-x1)(x-x2)....(x-x_{n-1})}{(x_n-x0)(x_n-x1)(x_n-x2)....(x_n-x_{n-2})}$$

It must be noted here that the factors in the numerator are exactly same as the denominator but only difference is that instead of x, x_0 is written. The same rule is applicable for all other values.

Though this method requires more calculations but can be applied to all types of problem of interpolation and extrapolation. To reduce computational work Newton's method could be used. The answer will be the same by both methods.

Illustration 2:

The following table gives the marks obtained by the students out of 10 in statistics. Estimate the number of students obtaining 4 marks.

Marks in statistics	0	2	3	5	6
No. of Students	5	7	8	10	12

(B.com, Kurukshetra 1997, Meerut 1996)

Solution :

Since the difference between independent variable is not same, we will apply lagrange's method.

Marks		**No. of Students**	
x		**y**	
0	x_0	5	y_0
2	x_1	7	y_1
3	x_2	8	y_2
5	x_3	10	y_3
6	x_4	12	y_4

$x = 4$, $y_x = ?$

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)(x_0-x_4)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_4)}$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)(x_2-x_4)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)(x_3-x_4)}$$

$$+ y_4 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_0)(x_4-x_1)(x_4-x_2)(x_4-x_3)}$$

$$y_x = \frac{5\times2\times1\times-1\times-2}{-2\times-3\times-5\times-6} + \frac{7\times4\times1\times-1\times-2}{2\times-1\times-3\times-4} + \frac{8\times4\times2\times-1\times-2}{3\times1\times-2\times-3} + \frac{10\times4\times2\times1\times-2}{5\times3\times2\times-1} + \frac{12\times4\times2\times1\times-1}{6\times4\times3\times1}$$

$$y_x = \frac{1}{9} - \frac{7}{3} + \frac{64}{9} + \frac{16}{3} - \frac{4}{3}$$

$$y_x = 0.1111 - 2.3333 + 7.1111 + 5.3333 - 1.3333$$

$$y_x = 12.5555 - 3.6666$$

$$y_x = 8.8889$$

$$y_x = 8.9$$

So, the number of students obtaining 4 marks are 9.

Illustration 3 :

Solve the illustration 1 by using lagrange method.

Solution :

x		**y**	
2	x_0	158	y_0
6	x_1	110	y_1
8	x_2	62	y_2
9	x_3	53	y_3

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}$$

$$y_x = 158 \frac{(5-6)(5-8)(5-9)}{(2-6)(2-8)(2-9)} + 110 \frac{(5-2)(5-8)(5-9)}{(6-2)(6-8)(6-9)}$$

$$+ 62 \frac{(5-2)(5-6)(5-9)}{(8-2)(8-6)(8-9)} + 53 \frac{(5-2)(5-6)(5-8)}{(9-2)(9-6)(9-8)}$$

$$y_x = 158 \times \frac{-1\times-3\times-4}{-4\times-6\times-7} + 110 \times \frac{3\times-3\times-4}{4\times-2\times-3} + 62 \times \frac{3\times-1\times-4}{6\times2\times-1} + 53 \times \frac{3\times-1\times-3}{7\times3\times1}$$

$$\frac{158}{14} + \frac{330}{2} - 62 + \frac{159}{7}$$

$$y_x = 137$$

Illustration : 4

Find out the percentage of terrorist upto age of 35 years.

Age	% of terrorist
Under 25 years	52.0
Under 30 years	67.3
Under 40 years	84.1
Under 50 years	94.4

(B.Com., Kashmir Univ., 1997)

Solution :

Since the differences between independent variable are not same, so we can use lagrange's method here.

Age		% of terrorist	
under 25 years	x_0	52.0	y_0
under 30 years	x_1	67.3	y_1
under 40 years	x_2	84.1	y_2
under 50 years	x_3	94.4	y_3

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}$$

Putting the values

$$y_{35} = 52 \frac{(35-30)(35-40)(35-50)}{(25-30)(25-40)(25-50)} + 67.3 \frac{(35-25)(35-40)(35-50)}{(30-25)(30-40)(30-50)}$$

$$+ 84.1 \frac{(35-25)(35-30)(35-50)}{(40-25)(40-30)(40-50)} + 94.4 \frac{(35-25)(35-30)(35-40)}{(50-25)(50-30)(50-40)}$$

$$y_{35} = 52 \frac{(5)(-5)(-15)}{(-5)(-15)(-25)} + 67.3 \frac{(10)(-5)(-15)}{(5)(-10)(-20)} + 84.4 \frac{(10)(5)(-15)}{(15)(10)(-10)} + 94.4 \frac{(10)(5)(-5)}{(25)(20)(10)}$$

$$y_{35} = 10.4 + 50.48 + 42.05 - 4.72$$

$$y_{35} = 77.44$$

Hence, 77.41% terrorists are under age of 35 years.

Illustration 5 :

Estimate the number of persons whose daily incomes are between Rs. 30 and Rs. 40

Daily income (In Rs.)	No. of persons.
15 - 20	73
20 - 30	97
30 - 45	110
45 - 55	180
55 - 70	140

(M.Com., Vikram, 1993)

Solution :

Since the class intervals are not same, we will apply the lagrange's method

Daily income in Rs.		No. of persons.
upto $20 x_0$	73	y_0
upto $30 x_1$	170	y_1
upto $45 x_2$	280	y_2
upto $55 x_3$	460	y_3
upto $70 x_4$	600	y_4

$x = 40$, $y_x = ?$

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)(x_0-x_4)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_4)}$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)(x_2-x_4)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)(x_3-x_4)}$$

$$+ y_4 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_0)(x_4-x_1)(x_4-x_2)(x_4-x_3)}$$

$$y_x = 73 \frac{(40-30)(40-45)(40-55)(40-70)}{(20-30)(20-45)(20-55)(20-70)} + 170 \frac{(40-20)(40-45)(40-55)(40-70)}{(30-20)(30-45)(30-55)(30-70)}$$

$$+ 280 \frac{(40-20)(40-30)(40-55)(40-70)}{(45-20)(45-30)(45-55)(45-70)} + 460 \frac{(40-20)(40-30)(40-45)(40-70)}{(55-20)(55-30)(55-45)(55-70)}$$

$$+ 600 \frac{(40-20)(40-30)(40-45)(40-55)}{(70-20)(70-30)(70-45)(70-55)}$$

$$y_x = 73 \frac{(10)\times(-15)\times(-30)}{(-10)\times(-25)\times(-35)\times(-50)} + 170 \frac{(20)\times(-5)\times(-15)\times(-30)}{(10)\times(-15)\times(-25)\times(-40)} + 280 \frac{(20)\times(10)\times(-15)\times(-30)}{(25)\times(15)\times(-10)\times(-25)}$$

$$+ 460 \frac{(20)\times(10)\times(-5)\times(-30)}{(35)\times(25)\times(10)\times(-15)} + 600 \frac{(20)\times(10)\times(-5)\times(-15)}{(50)\times(40)\times(25)\times(15)}$$

$y_x = -3.754 + 51 + 268.8 - 105.143 + 12 = 222.903$ or 223 persons

Number of persons whose daily income is upto Rs. 40 = 223

Number of persons whose daily income is upto Rs. 30 = 170

So Number of persons whose daily incomes is between

Rs. 30 - 40 = 223 – 170 = 53 persons.

Illustration : 6

From the following data, estimate the number of workers whose daily earning is Rs. 19 but not exceeding Rs. 25.

Income (Rs.)	No. of persons
1 and not exceeding 9	50
10 and not exceeding 19	70
19 and not exceeding 28	203
28 and not exceeding 37	406
37 and not exceeding 46	304

[M.A. Econ. Rajasthan Univ., B. Com., Kerala Univ., M. Com., Jabalpur Univ.]

Solutions.

Since the class intervals are not uniform throughout, we will apply the Lagrange's method. Estimate the number of persons getting income not exceeding Rs. 25.

Income (Rs.)		No. of persons	
Income not exceeding 9	x_0	50	y_0
Income not exceeding 19	x_1	120	y_1
Income not exceeding 28	x_2	323	y_2
Income not exceeding 37	x_3	729	y_3
Income not exceeding 46	x_4	1,033	y_0

Here, $x = 25$

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)(x_0-x_4)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_4)}$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)(x_2-x_4)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)(x_3-x_4)}$$

$$+ y_4 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_0)(x_4-x_1)(x_4-x_2)(x_4-x_3)}$$

$$y_{25} = 50 \frac{(25-19)\,(25-28)\,(25-37)\,(25-46)}{(9-19)\,(9-28)\,(9-37)\,(9-46)} + 120 \frac{(25-9)\,(25-28)\,(25-37)\,(25-46)}{(19-9)\,(19-28)\,(19-37)\,(19-46)}$$

$$+ 323 \frac{(25-9)\,(25-19)\,(25-37)\,(25-46)}{(28-9)\,(28-19)\,(28-37)\,(28-46)} + 729 \frac{(25-9)\,(25-19)\,(25-28)\,(25-46)}{(37-9)\,(37-19)\,(37-28)\,(37-46)}$$

$$+ 1033 \frac{(25-9)\,(25-19)\,(25-28)\,(25-37)}{(46-9)\,(46-19)\,(46-28)\,(46-37)}$$

$$= 50 \frac{(6)\,(-3)\,(-12)\,(-21)}{(-10)\,(-19)\,(-28)\,(-37)} + 120 \frac{(+16)\,(-3)\,(-12)\,(-21)}{(+10)\,(-9)\,(-18)\,(-27)} + 323 \frac{(+16)\,(+6)\,(-12)\,(-21)}{(+19)\,(+9)\,(-9)\,(-18)}$$

$$+ 729 \frac{(+16)\,(+6)\,(-3)\,(-21)}{(+28)\,(+18)\,(+9)\,(-9)} + 1033 \frac{(+16)\,(+6)\,(-3)\,(-12)}{(+37)\,(+27)\,(+18)\,(+9)}$$

$$= 1.15 + 33.19 + 282.07 - 108 + 22.06 = 230.47 \text{ or } 230 \text{ app.}$$

Number of persons whose income does not exceed Rs. 25 = 230.

Number of persons whose income does not exceed Rs. 19 = 120.

Hence the number of persons whose income is Rs. 19 but not exceeding Rs. 25 = (230 – 120) = 110.

Activity

1. Anubhooti has to write down the Newton's Divided Differences formula upto fourth leading divided difference. Help her in writing this.

   _____

   _____

2. Mauli has been given a question to solve by his teacher. He has been given a case of unequal intervals of x, 4 values are given and 1 has to be interpolated.
   Now he could not recall the formula. Help him in witting down the Lagrange's formula to be used.

   _____

   _____

16.4 Inverse Interpolation

In the previous and this unit, we have discussed various methods of interpolation and extrapolation. In previous pages we were given situation of interpolation where a set of values of x (independent variable) are given and we have to estimate y (dependent variable) = f(x) for a particular value of x.

Let us now observe the reverse situation.

For example, we are given a set of values of x and y. In this case we are interested to find the value of x for a specified value of y.

This situation is termed as 'Inverse Interpolation.

The following formula is used to calculate inverse interpolation. When four arguments a_0, a_1, a_2 and a_3 the value of x are given.

$$x = \frac{[f(x)-f(a_1)][f(x)-f(a_2)][f(x)-f(a_3)]}{[f(a_0)-f(a_1)][f(a_0)-f(a_2)][f(a_0)-f(a_3)]} \times a_0 + \frac{[f(x)-f(a_0)][f(x)-f(a_2)][f(x)-f(a_3)]}{[f(a_1)-f(a_0)][f(a_1)-f(a_2)][f(a_1)-f(a_3)]} \times a_1$$

$$+ \frac{[f(x)-f(a_0)][f(x)-f(a_1)][f(x)-f(a_3)]}{[f(a_2)-f(a_0)][f(a_2)-f(a_1)][f(a_2)-f(a_3)]} \times a_2 + \frac{[f(x)-f(a_0)][f(x)-f(a_1)][f(x)-f(a_2)]}{[f(a_3)-f(a_0)][f(a_3)-f(a_1)][f(a_3)-f(a_2)]} \times a_3$$

Illustration 7:

You are given the following information :

x	5	6	9	11
$f(x)$	12	13	14	16

Find the value of x when $f(x) = 15$.

Solution.

In the usual notation of Lagrange's formula :

x	$a_0 = 5$	$a_1 = 6$	$a_2 = 9$	$a_3 = 11$
$y = f(x)$	12	13	14	16

We have to calculate x when $f(x) = 15$. Using inverse interpolation formula :

$$x = \frac{(15-13)(15-14)(15-16)}{(12-13)(12-14)(12-15)} \times 5 + \frac{(15-12)(15-14)(15-16)}{(13-12)(13-14)(13-16)} \times 6$$

$$+ \frac{(15-12)(15-13)(15-16)}{(14-12)(14-13)(14-16)} \times 9 + \frac{(15-12)(15-13)(15-14)}{(16-12)(16-13)(16-14)} \times 11$$

$$= \frac{(2)(1)(-1)}{(-1)(-2)(-4)} \times 5 + \frac{(3)(1)(-1)}{(1)(-1)(-3)} \times 6 + \frac{(3)(2)(-1)}{(2)(1)(-2)} \times 9 + \frac{(3)(2)(1)}{(4)(3)(2)} \times 11$$

$$= 1.25 - 6 + 13.5 + 2.75 = 11.5$$

16.5 Parabolic Curve Method

Like lagrange's method, this method is also a universal method. Parabolic curve method is known as the method of simultaneous equations. This method can be used to solve all type of problems of interpolation. But in practice, this is used when the number of variables are less (say 3 or 4). When we use this method one variable is taken as independent variable and other is dependent variable. The independent variable will be x and dependent variable will be denoted by y, therefore, for a known x we can interpolate the value of y.

In this method, we have to solve a number of simultaneous equations to find the values. The number of equations to be solved is equal to the number of known values of x and y. The equation will be as follows.

$$y = a + bx + cx^2 + dx^3 \ldots\ldots + nx^n$$

This is a curve of the n^{th} order. The order for which equation is be raised depends upon the number of known values. The curve power will be less than 1, the number of known values.

For example, if the known values are three we, would use a curve of 2^{nd} order

To estimate the value of 'y' the following procedure are taken :

1. First of all, decide the power for which the parabolic curve will be applied. It is '1' less than the known values.

Parabolic Curve Equations

No. of known Values (n)	Powers of the Parabola (n – 1)	Equation
2	1	$y = a + bx$
3	2	$y = a + bx + cx^2$
4	3	$y = a + bx + cx^2 + dx^3$
5	4	$y = a + bx + cx^2 + dx^3 + ex^4$
n	n – 1	$y = a + bx + cx^2 + dx^3 + ex^4 + \ldots\ldots nx^{n-1}$

2. The independent variable 'x' for which dependent variable 'y' is to be estimated is taken as origin.

3. Then deviation from origin are taken for different given values of x.

4. After this, the parabolic curve equation are formed by putting the different values of y and deviation of x.

5. By solving all equations the value of 'a' is found out. This will be the value to be estimated.

Though this is a tedious method but has the advantage universal application. Now, parabolic curve method is not so popular because if the number of known values exceeds four, the calculations are more involved and hence time consuming.

Illustration 8 :

The following data are related to the cube of the values. Using parabolic curve method, find out the cube of 5.

Size	3	4	5	6	7
Cube Value	24	64	?	216	343

Solution :

Since the known values are four, we would fit a parabola of 3^{rd} order.

The equation would be,

$$y = a + bx + cx^2 + dx^3$$

We have to calculate the values of a, b, c and d.

Deviations of x from 5

x	– 2	– 1	0	+1	+2
y	27	64	y_0	216	343

Putting the values of x and y in the equation.

$27 = a - 2b + 4c - 8d$ .....(i)

$64 = a - b + c - d$ .....(ii)

$Y_0 = a$ .....(iii)

$216 = a + b + c + d$ .....(iv)

$343 = a + 2b + 4c + 8d$ .....(v)

By adding (ii) and (iv) eq.

$64 = a - b + c - d$

$216 = a + b + c + d$

$280 = 2a + 2c$ .....(vi)

By adding (i) and (v) eq.

$27 = a - 2b + 4c - 8d$

$343 = a + 2b + 4c + 8d$

$370 = 2a + 8c$ .....(vii)

Now, by solving (vi) and (vii) eq.

$280 = 2a + 2c$ .....(vi)

$370 = 2a + 8c$ .....(vii)

Multiplying (vi) eq. by 4

$1120 = 8a + 8c$

$370 = 2a + 8c$

– – –

$750 = 6a$

$a = 125$

Hence, the cube of 5 is 125.

Illustration 9 :

The following table presents the production of a firm for different years. Estimate the production for 1986 by using parabolic curve method.

Year	1971	1981	1991	2001
Production (In Tonnes)	36	44	50	60

Solution :

There are four known values, we fit a parabola of 3 rd order.

$y = a + bx + cx^2 + dx^3$

For determining the values of a, b, c and d, we will take deviations of x by taking 1986 as origin.

1986 = 0

We get,

	1971	1981	1986	1991	2001
x	-15	-5	0	5	15
y	36	44	y_0	50	60

By further simplification

x	–3	–1	0	+1	+3
y	36	44	y_0	50	60

Substituting the values of x and y.

We get equation,

$y = a + bx + cx^2 + dx^3$

$36 = a - 3b + 9c - 27d$ .....(i)

$44 = a - b + c - d$ .....(ii)

$y_0 = a$ .....(iii)

$50 = a + b + c + d$ .....(iv)

$60 = a + 3b + 9c + 27d$ .....(v)

Here, the value of y_0 is to be interpolated. So, we have to calculate the value of a for getting desired result.

By adding eq. (i) and (v)

$36 = a - 3b + 9c - 27d$

$60 = a + 3b + 9c + 27d$

$96 = 2a + 18c$ .....(vi)

adding eq. (ii) and (iv)

$44 = a - b + c - d$

$50 = a + b + c + d$

$94 = 2a + 2c$ .....(vii)

By solving eq. (vi) and (vii)

$96 = 2a + 18c$

$94 = 2a + 2c$

Multiplying eq. (vii) by 9

$96 = 2a + 18c$

$846 = 18a + 18c$

– – –

$-750 = -16a$

$16a = 750$

$a = 46.875$

Hence, the production for 1986 is 46.875 tonnes.

Illustration 10 :

The following table gives the data of the population of a city. Find out the population for 2001 using parabolic curve method.

Year	1995	1999	2003	2007
Population (In Lakhs)	50	56	68	90

Solution :

There are four known values, we would fit a parabola, of 3rd order.

By taking deviations of x.

Year	1995	1999	2001	2003	2007
Deviations x's	– 6	– 2	0	+2	6
y's	50	56	y_0	68	90

By further simplifications

x	– 3	– 1	0	1	3
y	50	56	y_0	68	90

The equation of 3rd order

$y = a + bx + cx^2 + dx^3$

By substituting the value of x we get,

$50 = a - 3b + 9c - 27d$ .....(i)

$56 = a - b + c - d$ .....(ii)

$y_0 = a$ .....(iii)

$68 = a + b + c + d$ .....(iv)

$90 = a + 3b + 9c + 27d$ .....(v)

By adding (i) and (v) equations

$50 = a - 3b + 9c - 27d$

$90 = a + 3b + 9c + 27d$

$140 = 2a + 18c$ .....(vi)

By adding (ii) and (iv) equations

$56 = a - b + c - d$

$68 = a + b + c + d$

$124 = 2a + 2c$ .....(vii)

By solving equation (vi) and (vii)

$140 = 2a + 18a$

$124 = 2a + 2c$

Multiplying equation (vii) by 9

$140 = 2a + 18c$

$1116 = 18a + 18c$

– – –

$-976 = -16a$

$16a = 976$

$a = 61$

Hence the population for 2001 is 61 lakhs.

Illustration 11 :

Using following data, estimate the value of U_2 by parabolic curve method.

x	1	2	3	4	5
U_x	7	–	13	21	37

(I.C.W.A.I., June 1997)

Solution :

Since only four values are known, we assume a 3rd degree polynomial. The equation for U_x.

$U_x = a + bx + cx^2 + dx^3$

Putting the values of x = 1, 3, 4, 5 successively.

We get,

$7 = a + b + c + d$ .....(i)

$13 = a + 3b + 9c + 27d$ .....(ii)

$21 = a + 4b + 16c + 64d$ .....(iii)

$37 = a + 5b + 25c + 125d$ .....(iv)

Subtracting equation (i) from (ii)

$$13 = a + 3b + 9c + 27d$$
$$7 = a + b + c + d$$
$$6 = 2b + 8c + 26d$$

or $3 = b + 4c + 13d$ .....(v)

Subtracting equation (ii) from (iii)

$$21 = a + 4b + 16c + 64d$$
$$13 = a + 3b + 9c + 27d$$
$$8 = b + 7c + 37d$$ .....(vi)

Subtracting equation (iii) from (iv)

$$37 = a + 5b + 25c + 125d$$
$$21 = a + 4b + 16c + 64d$$
$$16 = b + 9c + 61d$$ .....(vii)

Subtracting equation (v) from (vi)

$$8 = b + 7c + 37d$$
$$3 = b + 4c + 13d$$
$$5 = 3c + 24d$$ .....(viii)

Subtracting equation (vi) from (vii)

$$16 = b + 9c + 61d$$
$$8 = b + 7c + 37d$$
$$8 = 2c + 24d$$ .....(ix)

Subtracting (viii) from (ix)

$$8 = 2c + 24d$$
$$5 = 3c + 24d$$
$$+3 = -c$$
$$c = -3$$

Putting the value of $c = -3$ in equation (viii), we get,

$$5 = -9 + 24d$$
$$5 + 9 = 24d$$
$$24d = 14$$
$$d = \frac{14}{24} = \frac{7}{12}$$

Putting the values of c and d in equation (v)

$$3 = b + 4c + 13d$$
$$3 = b - 12 + 13 \times \frac{7}{12}$$
$$3 = b - 12 + \frac{91}{12}$$
$$3 + 12 - \frac{91}{12} = b$$
$$\frac{36 + 144 - 91}{12} = b$$
$$\frac{89}{12} = b$$

Putting the value of b, c and d in equation (i)

$$7 = a + b + c + d$$

$$7 = a + \frac{89}{12} - 3 + \frac{7}{12}$$

$$7 + 3 - \frac{7}{12} - \frac{89}{12} = a$$

$$10 - \frac{96}{12} = a$$

$$10 - 8 = a$$

$$a = 2$$

Therefore,

$$U_x = a + bx + cx^2 + dx^3$$

putting the values of a, b, c and d.

$$a = 2, \quad b = \frac{89}{12}, \quad c = -3, \quad d = \frac{7}{12}$$

when x = 2

$$U_2 = 2 + \frac{89}{12} \times 2 - 3 \times 2^2 + \frac{7}{12} \times 2^3$$

$$U_2 = 2 + \frac{89}{6} - 12 + \frac{14}{3}$$

$$U_2 = \frac{12 + 89 - 72 + 28}{6}$$

$$U_2 = \frac{57}{6} = \frac{19}{2} = 9.5$$

So the value of $U_2 = 9.5$

16.6 Summary

There may be some different situations where direct Binomial expansion method and Newton's advancing difference method are not applicable. Newton's divided difference method is used when the independent variable x increases by unequal intervals. In this method we subscript independent variable (x) and dependent variable (y) serially. After this a table of divided differences is constructed and the formula is applied to find out the value to be interpolated. Lagrange's method can be used whether the series increases by regular or irregular intervals and dependent variable (y) is in the beginning or at the end. In inverse interpolation we have to find out the value of x for a specified value of y. Although parabolic curve method can be used to solve all type of problems but it is generally used when the number of variables are less. A number of equations are solved to find the values. The number of equations to be solved is equal to the number of known values of x and y. It is a time consuming and a tedious method

16.7 Key Words

Newton's Divided Difference Method : The method which is used when the independent variable x increases by unequal intervals.

Lagrange's Method : The method which is used when series increases by regular or irregular interval and dependent variable is in the beginning or at the end.

Inverse Interpolation : The method which is used to find the value of x for a specified value of y.

Parabolic Method : The method which is used to solve all type of problems of interpolation and the number of variables are less.

16.8 Self Assessment Questions

1. Write down the Newton's divided differences formula upto third leading divided differences.
2. State the conditions under which lagrange's method is used?

3. From the following data, obtain the value of y when , x = 6 by using Newton's divided difference method.

x	3	5	7	8	10
y	180	154	120	110	90

(Ans. 135)

4. Using Newton's divided difference method, find out the premium payable at 15 years of age of a life insurance policy from the following data about the premium payable at different years :

Age (in years)	10	18	20	25
Premium (in Rs.)	130	250	310	360

(Ans. Rs. 161)

5. From the following table, Interpolate the missing figure using lagrange's method,

x	0	1	3	6
y	180	250	?	410

(B.Com., Meerut 1995)

(Ans. 352)

6. Using Lagrange's formula, estimate the production of a firm for the year 2003.

Year	2001	2002	20004
Production (in tonnes)	8.5	12	10

Modified (M.Com., Kanpur 1998)

(Ans. 12.5 tonnes)

7. Below are given the value of a function f(x) for certain values of x :

x	0	1	3	4
$f(x)$	5	6	50	105

Find $f(2)$, using lagrange's method. (I.C.W.A.I. Dec. 1975)

(Ans. 19)

8. From the following table, using lagrange's method interpolate the number of students securing first division marks. (Means equal to or above 120)

Marks (out of 200)	above 50	above 72	above 90	above 110	above 140
No. of students	150	130	90	50	20

(M.A. Meerut, 1997)

(Ans. above $y_{119} = 37$)

9. From the following table estimate the number of persons earning Rs. 35000, using lagrange's method.

Earnings (in thousand Rs.)	25	30	40	50
No. of persons	50	60	70	100

(B.com, Allahabad 1998)

(Ans. 65)

10. From the following information interpolate the number of workers earning between Rs. 6000 and Rs. 7000. Use lagrange's method.

Earnings (in 000 Rs.)	0-2	2-3	3-6	6-8
No. of workers	10	15	35	20

(M.A. Meerut 1999)

(Ans. $y_7 = 70$, 70-60 = 10)

11. From the following table fine out the value of x when y = 23, using inverse interpolation formula.

x	1	2	4
y	3	11	31

(Ans. 3.2857)

12. From the following table relating to annual earnings of a firm. Interpolate the earnings of the firm for the year 2007 by parabolic curve method.

Year	2005	2006	2007	2008
Earnings (in Rs. crore)	10	12	?	18

(Ans. 14.67 crores)

13. From the following data, using parabolic curve method interpolate the value of x = 2

x	0	1	3	4
y	10	12	100	210

(Ans 38)

14. By using parabolic curve method, estimate the most likely annuity at the age of 45 years.

Age (years)	40	50	60	70
Annuity (Rs.)	124	144	184	240

(Ans Rs. 131.25)

15. From the following, estimate the value of $\log_{10} 656$

$\log_{10} 654 = 2.8156$
$\log_{10} 658 = 2.8182$
$\log_{10} 659 = 2.8189$
$\log_{10} 661 = 2.8202$

use (a) parabolic curve method (b) lagrange's method.

(Ans. $\log_{10} 656 = 2.8168$)

16.9 Reference Books

1. Gupta, S.P. Statistical Methods.
2. Saha, S. Business Statistics.

Unit - 17 Association of Atributes

Structure of Unit:

17.0 Objectives

After completing this unit, you would be able to:

- Understand the qualitative aspects in statistics;
- Enhance your understanding about the topic 'Association of Attributes';
- Know about the various types of attributes;
- Test the consistency of given statistical data;
- Point out the various methods of determining association.

17.1 Introduction

In any study, the observations on units or individuals are of two types. Firstly, the observations may be about the measurement of certain characters of the units for example; income, expenditure, volume, height, weight, yields, etc. (numerical), Secondly, the observations may be for some characters or attributes of the units or respondents which they possess, for example the level of education, blindness, liking, environment etc. (descriptive).

Thus, we can say, characteristics possessed by an individual item may be classified into (1) numerical and (2) descriptive. The characteristics which are capable of being measured quantitatively are termed as statistics of variables (numerical classification). The characteristics which are not capable of quantitative measurement are termed as statistics of attributes (descriptive classification).

Often the need to know the relationship or the extent of association between two or more qualitative or quantitative characters arises. The method of obtaining association between two qualitative characters is obtained by using a statistical tool, between two or more attributes called "Association of attributes".

17.2 Classification of Attributes

On the basis of an attribute, an observation of the population can be studied. For example, when we study the literacy of a village, then the population of the village is divided into two classes-one class of people who are literate and the other class who are not literate or illiterate. When you study more than one attribute, there will be more than two classes. For example, take the attribute literacy along with employed.

Then, the classes will be literate, not literate, employed, not employed, literate not employed, literate employed, employed not literate and not literate not employed.

Here it is must to note that a clear definition must be laid down of the various attributes under study. It means an item finds place in only one class by a demarcation between two classes. For example, the students of fashion designing course are divided into two categories – tall and short. First we have to lay down a standard height; on the basis of this demarcation line, we categories the students as tall and short: those below the standard height are short and those above the standard height are tall.

17.3 Dichotomy and Manifold Classification

A classification of the simple kind considered in which each class is divided into two sub-classes, has been termed by logicians classification, or to use the more strictly applicable term, division by dichotomy (cutting in two). The classifications of most statistics are not dichotomous, for most usually a class is divided into more than two sub-classes, called manifold classification.

17.4 Positive and Negative Classes

The attributes may be positive or negative. If the attribute is present, it is termed as positive class; and its contrary or opposite is known as negative class. For theoretical purposes it is necessary to have some simple notation for the classes formed and for the numbers of observations assigned to each.

The positive class, in which the attribute is present, is denoted by capital letters, *A, B, C* etc. The negative class in which the attributes is absent is denoted by small Greek letters, α (alpha), β (Beta), γ (Gamma).

Thus, if *A* represents the attribute blindness, α represents sight (non-blindness); if *B* stands for deafness, β stands for hearing. Generally α is equivalent to not *A* or an object or individual not possessing the attribute *A*; the class a is equivalent to the class none of the members of which possesses the attribute *A*. *N* will denote the universe. The class frequencies are expressed by putting the symbols within the brackets, i.e., (*A*), (*B*), (AB). *N* denotes the number of numbers without any specification of attributes and is not placed in bracket.

N, A, B, AB etc. are positive classes

α, β etc. are negative classes

αB, $A\beta$, etc. are pairs of contrary classes

Activity A:

1. Is it necessary to study "Association of Attributes"? If yes, then give suitable reasons.

17.5 Number of Classes

The total number of classes comprising of the various attributes can be determined by 3^n, *n* representing the number of attributes. If one attribute is studied, then there will be $3^1 = 3$ classes. Thus, if literacy is studied, the presence of literacy is represented by *A*, its absence by α and total by *N*, then, there will be 3 classes; i.e., *A*, α and *N*.

If two attributes are studied, the number of classes will be $3^2 = 9$ classes; i.e.,

$(A), (\alpha), (B), (\beta), (AB), (A\beta), (\alpha\beta), (\alpha B), (N)$

For two attributes A and B

	A	***α***	***Total***
B	***(AB)***	***(αB)***	***(B)***
β	***(Aβ)***	***(αβ)***	***(β)***
Total	***(A)***	***(α)***	***N***

The above table is similar to a contingency table and hence certain relations are obvious

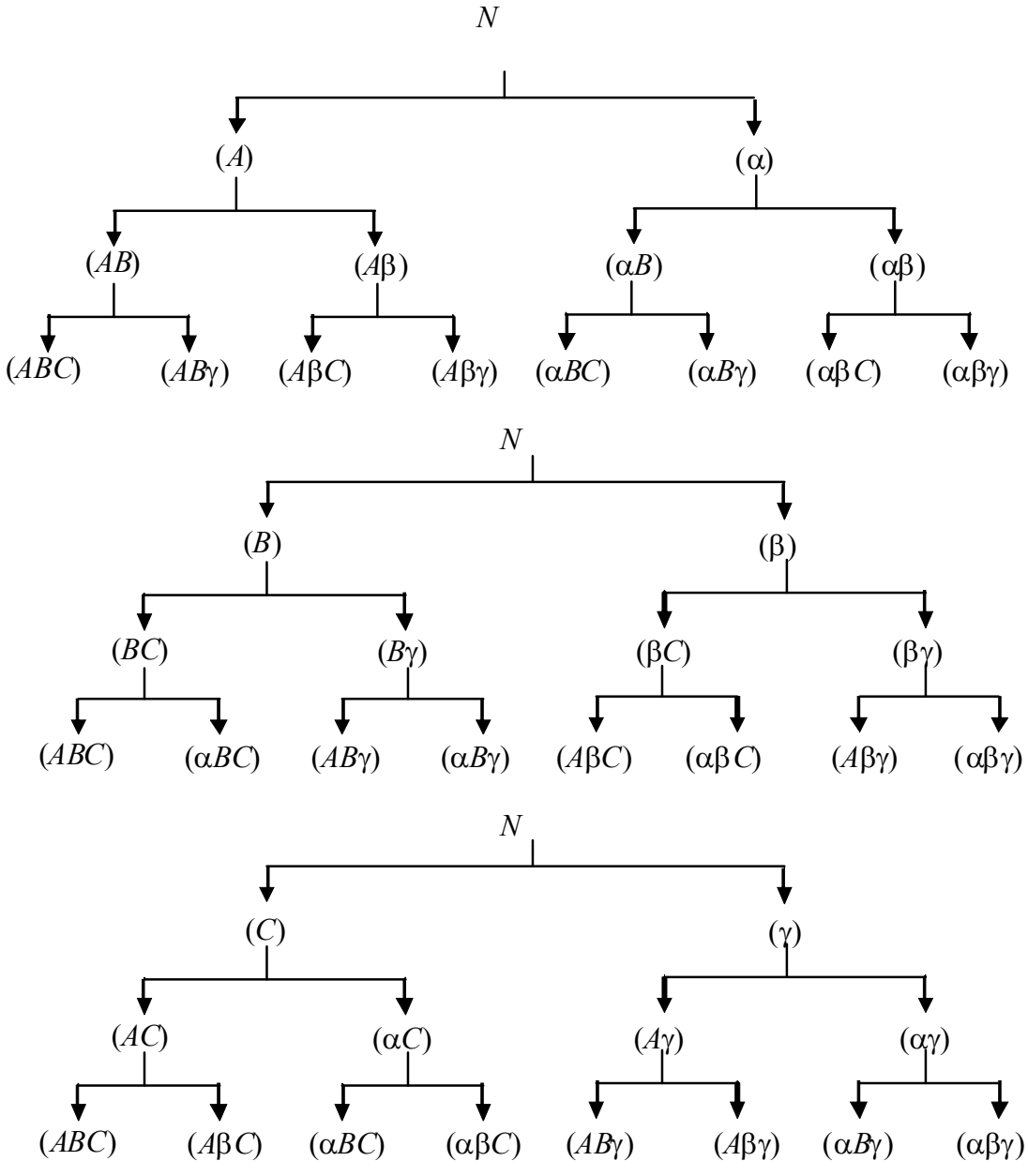
$(A) = (AB) + (A\beta)$
$(B) = (AB) + (\alpha B)$
$(\alpha) = (\alpha B) + (\alpha\beta)$
$(\beta) = (A\beta) + (\alpha\beta)$
$(AB) + (A\beta) + (\alpha B) + (\alpha\beta) = N$
$(A) + (\alpha) = (B) + (\beta) = N$

The above relations are helpful in finding out anyone of the missing class frequency.

If three attributes are studied, the number of classes will be $3^3 = 27$ classes.

	A		**α**		**Total**
	B	**β**	***B***	**β**	
C	***(ABC)***	**(*A*β*C*)**	**(α*BC*)**	**(αβ*C*)**	**(*C*)**
γ	**(*AB*γ)**	**(*A*β*γ*)**	**(α*B*γ)**	**(αβγ)**	**(γ)**
Total	**(*AB*)**	**(*A*β)**	**(α*B*)**	**(αβ)**	***N***

The totals and subtotals of all frequencies given in above table are depicted through the following chart.



Three attributes case:

$(A) = (ABC) + (A\beta C) + (AB\gamma) + (A\beta\gamma)$

$(\alpha) = (\alpha BC) + (\alpha\beta C) + (\alpha B\gamma) + (\alpha\beta\gamma)$

$(B) = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma)$

$(\beta) = (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$

$(C) = (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C)$

$(\gamma) = (AB\gamma) + (A\beta\gamma) + (\alpha B\gamma) + (\alpha\beta\gamma)$

Activity B:

1. First understand the above flow charts of three attributes, and then make it at your own.

17.6 Ultimate Class Frequencies

Those classes which specify the attributes of the highest order are known as the ultimate classes and their frequencies are known as the ultimate class frequencies. The process comes to a stop only when we reach the frequencies of the highest order.

The number of ultimate classes is represented by 2^n, where n stands for the number of attributes under study. Thus, for one attribute, there will be $2^1 = 2$ classes, for two attributes, the ultimate classes will be $2^2 = 4$; if three attributes, the ultimate classes will be $2^3 = 8$ and so on. For instance, when two attributes are studied, say, (A) and (B):

$(A) = (AB)(A\beta)$ These four are ultimate
$(\alpha) = (\alpha B)(\alpha\beta)$ class frequencies

Order of Classes

The order of classes depends upon the number of attributes under study. A class having one attribute is known as the class of the first order; a class having two attributes as the class of the second order and so on. N denotes the number of members without any specification of attributes as zero order and is not placed in brackets:

The following table gives the class frequencies of all orders and the total number of all class frequencies up to 3 attributes:

Order	*Attribute*	*Class Frequencies of Orders*	*No. in each order*	*Total No.*
0		N		1
0	A	N	1	
1		$(A)(\alpha)$	2	3
0	AB	N	1	
1		$(A)(B)(\alpha)(\beta)$	4	
2		$(AB)(A\beta)(aB)(\alpha\beta)$	4	9
0	ABC	N	1	
1		$(A)(B)(C)(\alpha)(\beta)(\gamma)$	6	
2		$(AB)(A\beta)(\alpha B)(\alpha\beta)$		
3		$(AC)(A\gamma)(\alpha C)(\alpha\gamma)$	12	
		$(BC)(B\gamma)(\beta C)(\beta\gamma)$		
		$(ABC)(AB\gamma)(A\beta C)(A\beta\gamma)$		27
		$(\alpha BC)(\alpha B\gamma)(\alpha\beta C)(\alpha\beta\gamma)$	8	

17.7 Frequency Determination

There are certain general rules for the determination of frequencies of various classes. The total number of observations is equal to the positive and negative frequencies of the same classes of the first order; for instance.

$$N = (A) + (\alpha)$$

Similarly

$$N = (B) + (\beta)$$

The frequencies can also be known with the help of a nine square table. Thus

	A	α	**Total**
B	(AB)	(αB)	(B)
β	$(A\beta)$	$(\alpha\beta)$	(β)
Total	(A)	(α)	N

Vertical totals

$(AB) + (A\beta) = (A)$

$(\alpha B) + (\alpha\beta) = (\alpha)$

$(B) + (\beta) = N$

Horizontal totals

$(AB) + (\alpha B) = (B)$

$(A\beta) + (\alpha\beta) = (\beta)$

$(A) + (\alpha) = N$

If known values are substituted for the symbols in the square, then the remaining values can be found out by addition or subtraction. Thus:

$$(A) - (AB) = (A\beta)$$

$$(\alpha) - (\alpha B) = (\alpha\beta)$$

$$N - (A) = (\alpha)$$

$$N - (B) = (\beta)$$

and so on. The following illustrations will clarify.

Illustration: 1

From the following ultimate class frequencies, find the frequencies of the positive and negative classes and the total number of observations:

$(AB) = 10$ $\quad$ $(A\beta) = 15$

$(\alpha B)) = 5$ $\quad$ $(\alpha\beta) = 40$

Solution:

It is required to find (A), (B), (α), β and N

$(A) = (AB) + (A\beta) = 10 + 15 = 25$

$(B) = (AB) + (\alpha B) = 10 + 5 = 15$

$(\alpha) = (\alpha B) + (\alpha\beta) = 5 + 40 = 45$

$(\beta) = (A\beta) + (\alpha\beta) = 15 + 40 = 55$

$N = (A) + (\alpha) = 25 + 45 = 70$

$N = (B) + (\beta) = 15 + 55 = 70$

The missing values of classes of the above illustration can also be found out with the help of the Nine Square Table; perhaps this is a convenient method when two attributes are under study.

Attribute	*A*	*α*	*Total*
B	*10* ***(AB)***	*5* ***(αB)***	*15* ***(B)***
β	*15* ***(Aβ)***	*40* ***(αβ)***	*55* ***(β)***
Total	*25* ***(A)***	*45* ***(α)***	***70*** ***N***

Illustration: 2

For two attributes A and B, we have $(AB) = 40$; $(A) = 60$; $N = 100$ and $(B) = 60$. Calculate the missing values.

Solution:

Given

	A	α	Total
B	40 (*AB*)	(α*B*)	60 (*B*)
β	(*A*β)	(αβ)	(β)
Total	60 (*A*)	(α)	100 *N*

Missing of Remaining Values

	A	α	
B		20 (60–40) (α*B*)	
β	20 (60–40) (*A*β)	20 (40–20) (αβ)	40 (100–60) (β)
		40 (100–60) (α)	

Activity C:

1. For two attributes A and B, we have (AB) = 50, (A) = 100, (N) = 200, (B) = 90. Calculate the missing values.

17.8 Test of Consistency

Class frequency can be positive or zero, but cannot be negative. Data observed may be described as consistent, if they do not conflict with each other. In case any class frequency is negative, then the given data are inconsistent. It is a simple test to be applied and verified whether the frequency or frequencies of classes are negative or not. If no conflict is there, no frequencies are negative, it is concluded that the given data are consistent.

Illustration: 3

Test for consistency, given $N = 100$, $(A) = 80$, (B) 60, $(AB) = 15$

Solution:

	(A)	(a)	
B	AB *(15)*	(a*B*) (45)	(B) (60)
β	(*A*β) (65)	(aβ) **-25** **therefore the given data are inconsistent**	(β) (40)
	(A) (80)	(a) (20)	N (100)

Activity D:

1. In a report on consumer's preference it was given that out of 500 surveyed 400 preferred variety A, 380 variety B and 270 persons were such who gave their likings for both the varieties. Is there any consistency in the data?

17.9 Types of Association

There are three types of association:

(A) Positive Association.
(B) Negative Association
(C) Independent Association

(A) Positive Association/Association: "When actual frequency is more than expected frequency, it is called positive association" i.e.

$$(AB) > \underset{\text{(Expected)}}{\frac{(A) \times (B)}{N}}$$
(Actual)

(B) Disassociation / Negative association: "when actual frequency is less than expected frequency, it is called negative association" i.e.

$$(AB) < \underset{\text{(Expected)}}{\frac{(A) \times (B)}{N}}$$
(Actual)

(C) Independent Association: "when actual frequency is equal to expected frequency, it is called independent association" i.e. i.e.,

$$(AB) = \underset{\text{(Expected)}}{\frac{(A) \times (B)}{N}}$$
(Actual)

Illustration: 4

Show from the following data whether (A) and (B) are independent, positively associated or negatively associated:

1. $N = 200$ $(A) = 40$ $(B) = 100$ $(AB) = 20$
2. $N = 400$ $(A) = 40$ $(B) = 160$ $(AB) = 25$
3. $N = 800$ $(A) = 168$ $(B) = 300$ $(AB) = 50$

Solution:

1. Expected frequency of $(AB) = \frac{(A) \times (B)}{N} = \frac{40 \times 100}{200} = 20$

   This is equal to the actual frequency of (AB), therefore (A) and (B) are independent.

2. Expected frequency of $(AB) = \frac{(A) \times (B)}{N} = \frac{40 \times 160}{400} = 16$

   Expected frequency is 16 Actual frequency of (AB) is 25, which is greater than the expected frequency; therefore, (A) and (B) are positively associated.

3. Expected frequency of $(AB) = \frac{(A)\text{x (B)}}{N} = \frac{168\text{x } 300}{800} = 63$

Actual frequency is 50, which is less than the expected frequency; therefore, (A) and (B) are negatively associated.

17.10 Methods of Determining Association

Association can be studied by any one of the following methods:

1. Comparison of Observed and Expected Frequencies.
2. Comparison of Proportions.
3. Yule's Coefficient of Association.
4. Yule's Coefficient of Colligation.
5. Pearson's Coefficient of Contingency.

1. Comparison of Observed and Expected Frequencies:

In this method the actual number of observation in compared with the expected frequencies. The expected frequency can be found by combination also. This will be clear from the following:

	Attribute	**Expected frequency**
(A) and (B)	(AB)	$\frac{(A)(B)}{N}$
(A) and (β)	(Aβ)	$\frac{(A)(\beta)}{N}$
(α) and (B)	(αB)	$\frac{(\alpha)(B)}{N}$
(α) and (β)	(αβ)	$\frac{(\alpha)(\beta)}{N}$

The limitation of this method is that, these methods only determine the nature of association, not the degree of association.

Illustration: 5

Can vaccination be regarded as a preventive measure for small pox from the data given below?

(i) Out of 2,000 persons in a locality exposed to Small - Pox, 500 in all were attacked.
(ii) If 2000 persons, 400 had been vaccinated; of these only 50 were attacked.

Solution:

Let (A): Vaccinated, (α) Not vaccinated
(B): Exempted from Small pox (β) Attack of Small Pox.

The missing values can be obtained from the following nine square table:

	***(A)* Vaccinated,**	**(α) Not vaccinated**	
(*B*) Exempted from Small pox	350 (*AB*)	1250 (α.)	1500 (B)
(β) Attack of Small Pox	**50** **(Aβ)**	450 (αβ)	**500** **(β)**
	400 **(A)**	1600 (α)	**2,000** **N**

Expected frequency of (AB) = $\frac{(A)x(B)}{N} = \frac{400\ x\ 1500}{2000} = 300$

(AB) > $\frac{(A)(B)}{N}$ = 350 (Actual) > 300 (Expected)

Therefore positively associated, Hence vaccination is a good preventive measure for small pox.

2. Comparison of Proportions:

Under this method, ratios or proportions of the concerned variables are compared. Relationship is given below:

Association	Proportion of B in A and α	Proportion of A in B and β
Independent	$\frac{(AB)}{A} = \frac{(\alpha B)}{(a)}$	$\frac{(AB)}{B} = \frac{(A\beta)}{\beta}$
Positive	$\frac{(AB)}{A} > \frac{(\alpha B)}{(a)}$	$\frac{(AB)}{B} > \frac{(A\beta)}{\beta}$
Negative	$\frac{(AB)}{A} < \frac{(\alpha B)}{(a)}$	$\frac{(AB)}{B} < \frac{(A\beta)}{\beta}$

Illustration: 6

Out of 800 literates in a particular district, the number of criminals was 10. Out of 9400 illiterates in the same district, the number of criminals was 200. On the basis of these figures do you find any association between illiteracy and criminality?

Solution:

Let A denote the attribute of illiteracy and B of criminality, α will denote literacy and non criminality.

Nine Square Table

	A	α	
B	200 (AB)	10 (αB)	210 (B)
β	9200 (Aβ)	790 (αβ)	9990 (β)
	9400 (A)	800 (α)	10200 (N)

Proportion of illiterate criminals to illiterate = $\frac{(AB)}{(A)} = \frac{200}{9400} = .0212\ or\ 2.12\%$

Proportion of literate criminals to literates: $\frac{(aB)}{(\alpha)} = \frac{10}{800} = .0125\ or = 1.25\%$

Hence the proportion of criminals is more in illiterates than in the literates, therefore criminality and illiteracy are positively associated.

3. Yule's Coefficient of Association:

The above mentioned methods will give us a rough idea about their association but the degree of association cannot be find out. Prof. Yule has suggested a formula to measure the association.

Yule's Coefficient of association = $Q = \frac{(AB)\ (\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

If Q ⟶ = 0 A and B are independent.

Q ⟶ = 1 A and B are positively associated

Q ⟶ = -1 A and B are negatively associated.

Illustration: 7

Investigate the association between darkness of eye colour in father and son from the following data using Yule's coefficient of association:

Fathers with dark eyes and sons with dark eyes	50
Fathers with dark eyes and sons without dark eyes	79
Fathers without dark eyes and sons with dark eyes	89
Fathers without dark eyes and sons without dark eyes	782

Solution:

Let (A) denote fathers with dark eyes.

Therefore, (α) will denote fathers without dark eyes.

Let (B) denote sons with dark eyes.

Therefore, (β) will denote sons without dark eyes.

We are given (AB) = 50, (Aβ) = 79, (αB) = 89, (αβ) = 782

$$Q = \frac{(AB)\ (\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Q = \frac{50\ x\ 782 - 79\ x\ 89}{50\ x\ 782 + 79\ x\ 89}$$

$$= \frac{39100 - 7031}{39100 + 7031}$$

$$= \frac{32069}{46131} = +.69\ Ans$$

Thus there is a high degree of positive association between the eye colours of father and son.

Illustration: 8

In an anti - malarial campaign in a certain area, quinine was administered to 812 persons out of a total population of 3248. The number of fever cases is shown below:

Treatment	*Fever*	*No Fever*
Quinine	20	792
No Quinine	220	2,216

Discuss the usefulness of quinine in checking malaria.

Solution:

	A	α	
B	20 (AB)	792 (αB)	812 (B)
β	220 (Aβ)	2216 (αβ)	2436 (β)
	240 (A)	3008 (α)	3248 N

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{20x2216 - 220x792}{20x2216 + 220x792}$$

$$= \frac{44{,}320 - 1{,}74{,}240}{44{,}320 + 1{,}74{,}240}$$

$$= \frac{-129920}{2{,}18{,}560} = \frac{-406}{683} = -.594$$

Thus, there is a negative association between treatment of quinine and attack of fever. Hence, quinine is useful is checking malaria.

4. Yule's Coefficient of Colligation:

This is another method for calculation of coefficient of association given by Yule, known as coefficient of colligation.

Formula:

$$\text{Co-efficient of Colligation (Y)} = \frac{1 - \sqrt{\frac{(A\beta)x\,(\alpha B)}{(AB)x\,(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)x\,(\alpha B)}{(AB)x\,(\alpha\beta)}}}$$

$$\text{Co-efficient of Association (Q)} = \frac{2y}{1 + y^2}$$

Illustration: 9

In a sample study and deafness in one locality, the following figures were obtained:

	Sane	***Insane***	***Total***
Deaf	20	40	60
Non-Deaf	50	25	75
	70	65	135

Trace the intensity of association between sanity and deafness. Also calculate the coefficient of colligation and derive Yule's Q theorem:

Solution:

(AB) = 20 (Aβ) = 50

(αB) = 40 (α β) = 25

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{20x25 - 50x40}{20x25 + 50x40}$$

$$= \frac{1500}{2500} = -.6$$

$$y = \frac{1 - \sqrt{\frac{(A\beta)\ (\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

$$= \frac{1 - \sqrt{\frac{50x40}{20x25}}}{1 + \sqrt{\frac{50x40}{20x25}}}$$

$= \frac{1-\sqrt{4}}{1+\sqrt{4}} = -\frac{1}{3} = -.333$

Co-efficient of Association $Q = \frac{2Y}{1+Y^2} = \frac{2(-.33)}{1+(-.33)^2}$

$= \frac{-.6667}{1.1089} = -.6$ *Verified.*

Illustration: 10

1660 candidates appeared for a competitive examination. 422 were successful, 256 had attended a coaching class; and of these 150 came out successful. Estimate the utility of the coaching classes.

Solution:

Let (*A*) = successful candidates, (α) = failure, (B) = attended coaching class; (β) did not attend coaching class.

	A	α	
B	150 (AB)	(αB)	256 (B)
β	(Aβ)	(αβ)	(β)
	422 (A)	(α)	1660 (N)

(αB) = (B) - (AB)

= 256-150 =106

(Aβ) = (A) - (AB)

= 422-150 = 272

(β) = (N) - (B)

= 1660-256 = 1404

(αβ) = (β) - (Aβ)

= 1404 - 272 = 1132

(α) = (αB) + (aβ)

= 106 + 1132 = 1238

	A	α	
B	150 (AB)	106 (αB)	256 (B)
β	272 (Aβ)	1132 (αβ)	1404 (β)
	422 (A)	1238 (α)	1660 (N)

$$Q = \frac{150\ x\ 1132 - 106\ x\ 272}{150x1132 + 160x272}$$

$$= \frac{1,69,800-28.832}{1,69,800+28,832} = \frac{1,40,968}{1,98,632} = 0.71$$

Thus, there is a positive association between successful candidates and those who attended coaching class.

5. Pearson's Coefficient of Contingency:

We have so far discussed dichotomous classification. Classification of data can be either dichotomous or manifold. When, the universe is divided into two groups, say, "rich" and "not rich" - "A" and "not A" but as A_1, A_2, A_3 etc. Similarly an another attribute, say B can be subdivided into B_1, B_2, B_3 etc. The frequency falling within the different classes can be arranged in the form of a contingency table.

	Attribute *A*						
Attribute B	*A* / *B*	A_1	A_2	A_3	...	A_n	Total
	B_1	$(A_1\ B_1)$	$(A_2\ B_1)$	$(A_3\ B_1)$	..	$(A_n\ B_1)$	(B_1)
	B_2	$(A_1\ B_2)$	$A_2\ B_2)$	$(A_3\ B_2)$	...	$(A_n\ B_2)$	(B_2)
	B_3	$(A_1\ B_3)$	$(A_2\ B_3)$	$(A_3\ B_3)$	...	$(A_n\ B_3)$	(B_3)
	...	...	...	...	...	...	
	B_n	$(A_1\ B_n)$	$(A_2\ B_n)$	$(A_3\ B_n)$	...		(B_n)
	Total	(A_1)	(A_2)	(A_3)	...	(A_n)	*N*.

A_1, A_2, A_3 etc. and B_1, B_2, B_3 etc. are the first order of the frequencies. And the frequencies of various cells are the frequencies of the second order. The total of A_1, A_2, A_3 etc. or the total of B_1, B_2, B_3, etc. would give grand total i.e. *N*.

The coefficient of mean square contingency or C according to Karl Pearson is:

$$C = \sqrt{\frac{x^2}{N + x^2}}$$

Where $x^2 = \sum \frac{(O-E)^2}{E}$

Where O = Observed Value

E = Expected Value

Illustration: 11

The following table shows the association among 1000 criminals between their weight and mentality. Calculate the coefficient of contingency between the two:

	Weight in Pounds					
Mentality	*90-120*	*120-130*	*130-140*	*140-150*	*150 and above*	*Total*
Normal	*50*	*102*	*198*	*210*	*240*	*800*
Weak	*30*	*38*	*72*	*30*	*30*	*200*
Total	*80*	*140*	*270*	*240*	*270*	*1000*

Solution:

Calculation of X^2 and c

Mentality	**Weight in Pounds**					
	90-120	*120-130*	*130-140*	*140-150*	*150 and above*	*Total*
Normal	50 **64**	102 **112**	198 **216**	210 **192**	240 **216**	800
Weak	30 **16**	38 **28**	72 **54**	30 **48**	30 **54**	200
Total	80	140	270	240	270	1000

$$x^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(50-64)^2}{64} + \frac{(102-112)^2}{112} + \frac{(198-216)^2}{216} + \frac{(210-192)^2}{192} + \frac{(240-216)^2}{216} + \frac{(30-16)^2}{16}$$

$$+ \frac{(38-28)^2}{28} + \frac{(72-54)^2}{54} + \frac{(30-48)^2}{48} + \frac{(30-54)^2}{54}$$

$$= \frac{196}{64} + \frac{100}{112} + \frac{324}{216} + \frac{324}{192} + \frac{576}{216} + \frac{196}{16} + \frac{100}{28} + \frac{324}{54} + \frac{324}{48} + \frac{576}{54}$$

$$= 3.06+.89+1.5+1.68+2.7+12.25+3.57+6+6.75+10.66$$

$$= 49.06 \text{ app.}$$

$$c = \sqrt{\frac{X^2}{N+X^2}} = \sqrt{\frac{49.06}{1000+49.06}} = .216 \text{ app. } \textit{Ans.}$$

Illustration: 12

Suppose that in a public opinion survey answers to the questions:

(a) Do you drink?
(b) Are you in favour of local option on sale of liquor?

Question (b)	***Question (a)***		
	Yes	***No***	***Total***
Yes	*56*	*31*	*87*
No	*18*	*6*	*24*
Total	*74*	*37*	*111*

Calculate coefficient of contingency between the two.

Solution:

O	E
56	$\frac{74x87}{111} = 58$
18	$\frac{74x24}{111} = 16$
31	$\frac{87x37}{111} = 29$
6	$\frac{37x24}{111} = 8$

$X^2 = \sum \frac{(O-E)^2}{E}$

$X^2 = \frac{4}{58} + \frac{4}{16} + \frac{4}{29} + \frac{4}{8}$

= .069 + .25 + .14 + .5

= .959 = .96 app.

$c = \sqrt{\frac{X^2}{N+X^2}} = \sqrt{\frac{.96}{111+.96}} = .0925$

Activity E:

1. Compare all the above five methods and describe which method is more appropriate and why in your opinion?

Association of Three Attributes (For Detail see 17.5)

Illustration: 13

In girls High School, there were 200 students. Their results in the quarterly, half-yearly and annual examination were as follows:

80 passed the quarterly examination.
75 passed the half-yearly examination.
96 passed the annual examination.
25 passed all three
46 failed in all three
29 passed the first two and failed in annual examination
42 failed in the first two but passed at least two examinations.

Solution:

Denoting:

Success in quarterly examination by (A) failure by (α)

Success in half-year examination by (B) and failure by (β)

Success in annual examination by (C) and failure by (*y*)

Therefore, the given values are reduced to:

N = 200; (ABC) = 25; (A) = 80; (αβy) = 46

(B) = 75; (ABY) = 29; (C) = 96; (αβC) = 42

We have to find out the value of:

$(ABC) + (AB\gamma) + (A\beta C) + (\alpha BC)$

Now, $(\alpha BC) + (A\beta C) + (ABC) + (\alpha\beta C) = (C)$

$\therefore (\alpha BC) + (A\beta C) = (C) - (ABC) - (\alpha\beta C) = 96\text{-}25\text{-}42 = 29$

$(\alpha BC) + (A\beta C) + (AB\gamma) + (ABC) = 29+29+25 = 83$

Thus the number of students who passed at least two examinations is 83.

17.11 Summary

Often the need to know the relationship or the extent of association between two or more qualitative or quantitative characters arises. The method of obtaining association between two qualitative characters is obtained by using a statistical tool, between two or more attributes called "Association of attributes".

A classification of the simple kind considered in which each class is divided into two sub-classes, has been termed by logicians classification, or to use the more strictly applicable term, division by dichotomy (cutting in two). The classifications of most statistics are not dichotomous, for most usually a class is divided into more than two sub-classes, called manifold classification.

The attributes may be positive or negative. If the attribute is present, it is termed as positive class; and its contrary or opposite is known as negative class. The positive class, in which the attribute is present, is denoted by capital letters, *A, B, C* etc. The negative class in which the attributes is absent is denoted by small Greek letters, α (alpha), β (Beta), γ (Gamma). Those classes which specify the attributes of the highest order are known as the ultimate classes and their frequencies are known as the ultimate class frequencies. Class frequency can be positive or zero, but cannot be negative. If no conflict is there, no frequencies are negative, it is concluded that the given data are consistent.

There are three types of association: positive association, negative association, independent association. Association can be studied by comparison of observed and expected frequencies, comparison of proportions, Yule's coefficient of association, Yule's coefficient of colligation, and Pearson's coefficient of contingency.

17.12 Self Assessment Questions

1. How would you distinguish between 'Association' and 'Correlation' as the terms used in Statistics.
2. What do you understand by Association of Attributes? How is its existence or non-existence determined? What is Partial Association?
3. What do you understand by 'Association of Attributes'? Discuss the methods by which it is measured.
4. How would you study association in contingency tables? What is meant by X^2 test? Explain fully.
5. When are two attributes said to be
   (a) Independent,
   (b) Positively associated, and
   (c) Negatively associated?
6. Examine the consistency of data when

   (i) N = 1000, (A) = 600, (B) = 50, (AB) = 50

   (Ans. Not consistent)

   (ii) N=2100, (A) = 1000, (B) = 1300, (AB) = 1100

   (Ans. Not consistent)

7. Show whether the attributes (A) and (B) are positively associated, negatively associated or independent in the following cases:

(i) N = 400, (A) = 200, (B) = 100, (AB) = 50

(ii) N = 500, (A) = 100, (B) = 200, (AB) = 60

(iii) N = 50, (A) = 100, (B) = 200, (AB) = 30

(Ans. (i) Independent (ii) positively associated (iii) negatively associated)

8. From the data given below, calculate Yule's coefficient of association between weight of children and their economic condition, and interpret it.

	Poor Children	Rich Children
Below Normal weight	75	23
Above Normal weight	5	42

(Ans. + 0.93: a high degree of positive association)

9. Find whether A and B are independent in the following case: (AB) = 256; (αB) = 768 (Aβ) = 48; ($\alpha\beta$) = 144

Ans. Independent

10. 1660 candidates appeared for a competitive examination, 422 were successful, 256 had attended a coaching class and 150 came out successful. Examine the utility of the coaching class with the help of Yule's Coefficient of Association

(Ans. Q = + .71)

11. Find out the co-efficient of association between the type of college training and success in teaching from the following table:

Institution	Successful	Unsuccessful	Total
Teacher's College	58	42	100
University	49	51	100
Total	107	93	200

(Ans. Q = - .18)

12. From the data given below test whether where is association between economic status and economic achievement:

	Rich	Poor
Educated	508	1559
Uneducated	905	1114

(Ans. Q = - 0.43)

13. Eighty eight residents of an Indian city, who were interviewed during a sample survey are classified below according to their smoking and tea drinking habits. Calculate Yule's coefficient of association and comment on its value.

	Smokers	Non-Smokers
Tea Drinkers	40	33
Non Tea Drinkers	3	12

(Ans. Q = + .66 app.)

14. In a group of 800 students, the number of married is 320. Out of 240 students who failed, 96 belonged to the married group. Find out whether the attributes marriage and failure are independent.

(Ans. Q = O)

15. Calculate the coefficient of association between intelligence in father and son from the following data -

Intelligent fathers with intelligent sons	248
Intelligent father with dull sons	81
Dull fathers with intelligent sons	92
Dull fathers with dull sons	579

(Ans. Q = +.9)

17.13 Reference Books

1. Richard I. Levin and David S. Rubin, Statistics for Management
2. Gupta, S. P., Statistical Methods
3. Yadav, Jain, Mittal, Statistical Methods.
4. Nagar, K. N., Statistical Methods.
5. Gupta, C.B. and Gupta, Vijay, An Introduction to Statistical Methods.

Unit - 18 Fundamentals of Probability

Structure of Unit:

18.0 Objectives

After completing this unit you would be able to:

- Define Probability
- Understand different approaches of Probability.
- Use counting techniques in calculations.
- Aware of different types of events which are related to Probability theorems for calculating Probability.
- Assess the importance and limitations of Probability.

18.1 Introduction

In our daily life, we use the term 'chances' or 'possibility', for example chances of heavy rainfall are very high today, chances of minimum temperature shooting up are very low, possibility of Indian team's victory in the forthcoming match is about 40 percent. These types of comments are passed due to some past experience or some guide or information related to these are available from any source. Thus, we express the chance of happening or not happening of a particular event on the basis of whatever relevant data available with us. If the chances are expressed in numerical form with statistical base, it is called 'Probability'.

18.2 Origin and Development of Probability

The credit for origin and development of probability goes to European gamblers of 17th century. In 1654

a French Person Chevalier De Mere presented his problem of throwing the dice and chance of getting a particular number before his friend and popular mathematician Blaise pascal. He solved this problem with consultation from another mathematician Pierre De Fermat. Thus Pascal and De Fermat gave the systematic and scientific foundation of mathematical theory of probability. Other mathematicians Huygens(in 1657), James Bernoulli(in 1713), De Moivre(in 1718) and Thomas Bayes(in 1764) developed different theorems of probability. Later Laplace, Gauss, Ronald Fisher, J.Neyman, Cheby Chev, A.Markov, A.N.Kolmogarov etc. made important contribution in expanding the theory of probability.

18.3 Definition and Approaches of Probability

'Probability' has been defined in three ways(approaches):-

18.3.1 Classical Approach

This concept was originated in 18th century. According to Laplace- "Probability is the ratio of the number of 'favourable' cases to the total number of equally likely cases". If an event can happen in 'm' ways, the probability of the event to occur will be

$$p = \frac{n}{m+n} \text{ or } \frac{\text{Number of favourable cases}}{\text{Number of total cases}}$$

The probability that the event will not occur shall be,

$$q = \frac{n}{m+n} \text{ or } 1\text{-}p \text{ or}$$

$$q = 1 - \frac{\text{Number of favourable cases}}{\text{Number of total cases}}$$

The main assumptions of this approach are as under-

(i) Probability of happening and non happening of an event can not exceed one or $p + q = 1$.

(ii)The outcomes of a random experiment are "equally likely" and "mutually exclusive".

Illustration1:

If a card is drawn at random out of a pack of cards:

(i) What it the Probability of drawing a queen?
(ii) What is the Probability of not drawing a Queen?

Solution :

Total number of cases = 52

Number of queens = 4, hence

Number of favourable cases = 4

$$\text{(i) Probability} = \frac{\text{Number of favourable cases}}{\text{Number of total cases}}$$

$$p = \frac{4}{52} \text{ or } \frac{1}{13}$$

(ii) Probability of not drawing a queen -

$$q = 1 - p \text{ or } 1 - \frac{4}{52} \text{ or } \frac{48}{52}$$

Limitations -

(1) In probability, 'equally likely events' means the events which have equal chances of happening or are equally probable. At times, events are not equally likely.

(2) If the outcomes of the events are biased or based on some trick, the theorems of probability can not be cannot be applied.

(3) If an event has infinite outcomes or is improbable, it is not possible to find the probability of this event.

(4) Probability is based not only on logic but it can be determined by experiment, experience or by actual observation.

18.3.2 Modern Approach

In this approach probability of happening or not happening of an event is found out on the basis of past experience or relative frequency is used. This theory is also called Statistical or Empirical Approach. According to Von Mises -"If the experiment be repeated a large number of times under essentially identical conditions, the limiting value of the ratio of the number of times of event A happens to the total number of trials increases indefinitely, is called the probability of the occurrence of A".

In the words of Croxton and Cowden, "Probability is the limit of the relative frequency of successes in infinite sequences of trials". If an event 'A' happens, 'm' number of times out of 'n' times, the relative frequency would be $\frac{m}{n}$. If n turns to be infinite then the probability of A =

$$p(A) = \text{limit } \frac{m}{n} \text{ where } n = \infty$$

The main assumption of this method is that if the experiments conducted upto infinity, its limit shall be on the basis of relative frequency and the experiments are of large number and random.

Illustration 2: Out of 2000 articles produced in a factory, 160 were found to be defective. Calculate the probability of defective articles in the production of the factory.

Solution

$$p(A) = \text{limit}_{n \to \infty} \frac{m}{n}$$

$$= \frac{\text{No. of defective articles}}{\text{Total no. of articles produced}} = \frac{160}{2000} = .08$$

Limitations-

(i) The infinite number of tests are far from reality.

(ii) The statistical measurement is based on past experience or present expectations which is not accurate all the time.

(iii) If the tests are conducted in different conditions, the outcomes will be different.

18.3.3 Subjective Approach

This latest approach was introduced by Frank Ramsey in his book, "The foundation of Mathematics and other Logical Essays(1926)". This approach says that we can calculate probability on the basis of whatever data or evidence or proof available with us. Thus, probability is totally based on subjectivity or individuality. In other words, probability depends on a person's personal beliefs, who is making probability statement.

Though this approach is very flexible and broad yet one has to be very careful otherwise probability assigned will be misleading or far away from actual happenings.

18.4 Counting Techniques

In the study of probability theory, the chance of occurrence of an event has to be assessed by counting the favourable outcomes to the happenings of the event from amongst all possible outcomes.

When the number of total events is comparatively small we can list them and calculate probability easily. This task becomes cumbersome when the number of possible outcomes are large. some counting methods have been developed to do calculation easily.

18.4.1 Factorial

Products of consecutive integers are included in many problems with permutations and other kinds of arrangements. For example one can write 4.3.2.1 or as $\lfloor 4$ or 4! and refer to it as "4 factorial".

Illustration- 3 : Compute the factorial of the following:

(i) 5! (ii) $\frac{10!}{6!\,4!}$

Solution :

(i) $5! = 5 \times 4 \times 3 \times 2 \times 1 = 5 \times 4! = 120$

(ii) $\frac{10!}{4!} = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(6 \times 5 \times 4 \times 3 \times 2 \times 1)(4 \times 3 \times 2 \times 1)}$

$= \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1}$

$= 210$

18.4.2 Permutations

Permutation refers to the specific arrangements of objects in a set where all elements are different and distinguishable. These arrangements are to be done without repetition of any individual object and all things should be different. It an event can happen in m ways and another event can happen in n ways, then the number of ways in which both the events can happen together will be m x n.

(a) Mathematical formula for permutation of 'r' things taken out of 'n' things is-

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

Illustration- 4 : Find the value of $8p_2$ and $4p_4$.

Solution

(i) $8p_2$ here $n = 8, r = 2$

$= \frac{8!}{(8-2)!}$ or $\frac{8!}{6!}$ or $\frac{8 \times 7 \times 6!}{6!}$

$= 8 \times 7 = 56$

(i) $4p_4 = \frac{4!}{(4-4)!}$ or $\frac{4!}{0!}$ or $\frac{4}{1!} = \frac{4 \times 3 \times 2 \times 1}{1} = 24$

(b) Number of Permutations when all things are not different. Suppose out of 'n' things 'p' things are same, 'q' are also alike and 'r' are also same, then the number of total permutations shall be:-

$$\frac{n!}{p!q!r!}$$

(c) Permutations of 'r' objects out of 'n' different objects when particular one object is always taken.

$${}^{r}P_1 \times {}^{n-1}P_{r-1}$$

(d) Permutations of excluding 'p' particular item then remaining items on excluding 'p' will be n - p and we have to fill 'r' places in the following way.

$${}^{n-p}P_r$$

(e**) Circular Permutation-**

(i) If clockwise and anticlockwise orders can be distinguished then the total number of desired permutations shall be (n - 1)!(ii) If clockwise and anticlockwise outcomes can not be distinguished then the total number of desired permutations shall be $\frac{1}{2}$ (n - 1) !

Activity-A

(A) Five routes are available from Delhi to Jaipur. In how may ways a person can go from Delhi to Jaipur and return if he can not return from the same route.
(B) How many permutations can be made from the word INDIA.
(C) In how many ways 4 gentlemen and 4 ladies can be seated around a round table provided no two gentlemen can sit together?
(D) In how many ways 11 players can be selected out of 15 players if 4 particular players are (i) always taken (ii) never taken.
(E) In how many ways 6 persons can sit on nine seats?

18.4.3 Combinations

In permutation the order or arrangement of the items is given importance, but in combination, order or arrangement is immaterial, e.g.' AB and BA are same in combination. Thus, selection of r things out of n given things, without any consideration of order, is called combination. For example, out of the three letters ABC, combinations of two letters at a time will be AB, AC or BC.

In the form of formula, $C\,\mathbf{2}=\frac{n!}{(n-r)!r!}$ or

$${}^{n}C_r = \frac{{}^{n}P_r}{r!}$$

Combinations with restrictions-

Number of combinations of 'r' things taken at a time out of 'n' different things if-

(i) p things are always included- then from n - p things now r - p things shall remain for selection. Therefore the number of combinations will be ${}^{n-p}C_{r-p}$

(ii)'p' things are never taken, then only n - pcr things are left and we have to take 'r' things from 'n'things. Therefore, the number of combinations will be ${}^{n-p}C_{r.}$

(iii) If some or all things are taken at a time out of 'n' different things. In this case each item can be taken into the combination as well as it can be left also. Thus, there are only two ways of taking or not taking an

item into a combination. In the same way there are only two ways of taking or not taking the second item and so on. Thus, number of total ways of n things = 2 x2x2 .....n or 2^n, but in the above ways there is one way in which one thing should be excluded. Therefore the number of total ways shall be = 2^n -1.

Illustration 5: In how many ways can a person give party to his 5 friends by inviting one or more at a time-

Solution-

He can invite 1,2,3,4,5 friends at a time. Therefore the number of total combinations will be 2^n -1 or 2^5-1 or 32-1 = 31 ways.

Activity-B

(i) What is the probability of drawing a heart or a king from a pack of cards?
(ii) Thirty balls are serially numbered and placed in a bag. Find the chance that the first ball drawn in a multiple of 5 or 6.
(iii) A drawer contains 8 pairs of black socks, 4 pairs of brown socks, 6 pairs of white socks and 5 pair of blue socks. If one pair is taken out from the drawer, what is the probability of-
(a) Getting a brown pair;
(b) Getting a white or blue pair;
(c) Not getting a black pair.
(d) Getting a black pair or a brown pair or a white pair or a blue pair?

Probability scale

The probability of happening an event may run from zero to one. If the probability is zero (p=0), it means that there is no possibility of happening of that particular event, or in other words, the event is impossible. If the probability is one (p=1), it means happening of that particular event is certain. The probability of happening an event cannot be negative.

18.5 Types of Events

For calculating probability it is necessary to understand the term Event, An event means the set of all possible outcomes of an experiments which may be of following types:-

1) **Simple and Compound events:** In simple events, we find the probability of happening or not happening of only one event, for example, getting 5 in a throw of a dice, drawing an ace from a pack of 52 cards, drawing a white ball from a bag containing 6 white balls and 4 red balls etc.
Compound event includes two or more events to happen simultaneously or one after the other.
For example, to draw an ace and a queen from a pack of cards, two coins are tossed simultaneously or one coin is tossed two times, these shall be called compound events.

2) **Equally likely events:** Such an event which has equal chance of happening. For example getting head or tail on tossing a coin is $\frac{1}{2}$, in the same way getting two in a throw of dice is $\frac{1}{6}$.

3) **Mutually exclusive events:** If two events cannot occur simultaneously, these are called mutually exclusive events. According to Spiegel "Two or more events are mutually exclusive if the occurrence of any one of them excludes the occurrence of the others." For example, in a throw of dice if 2 comes up, the chance of having other number ends.

3) **Dependent events:** If the happening of one event affects the happening of some other event, the two events are said to be dependent events. For example, in a pack of 52 cards there are 4 aces, if one card is drawn, probability of its being a king will be $\frac{4}{52}$, if the first drawn card is a king and it is not replaced,

then on drawing a card again, the probability of getting a king will now be $\frac{3}{51}$. The result of first event affects the result of second event, so these are called dependent events.

4) **Independent Events**: When happening of a certain event doesn't affect the happening or not happening of another, these events are known as independent events. For example, if one coin is tossed twice, event of tossing the coin for the first time doesn't affect the event of tossing the coin for the second time. similarly, probability of drawing a red card from a pack of 52 cards is $\frac{26}{52}$ or $\frac{1}{2}$. If the first drawn card is replaced and again one card is drawn, still the probability of drawing a red card will be $\frac{26}{52}$. Thus, drawing one card at a time twice (with replacement) are independent events. If more than two events are to happen, none of these affects the outcome of another events, then these all are independent events.

5) **Exhaustive event:** Events are said to be exhaustive when all the possible outcomes of random experiments are taken together. For example, while tossing a dice, the possible outcomes are 1,2,3,4,5 and 6 and thus the exhaustive number of cases is 6. Similarly for a throw of 3 coins, exhaustive number of cases will be 2 x 2 x 2=8 and for n coin it will be 2^n.

6) **Complementary events:** In an experiment if two events are mutually exclusive and exhaustive and sum of their probability is one, happening or not happening of one event is complementary to another, for example on tossing a coin, the event of getting a tail is complementary to the event of getting a head.

7) **Sample Space-** All the possible outcomes or events of a given experiment is called the sample space. it can be finite or infinite.

18.6 Probability Theorems

There are four important theorems of probability, namely;

1. Addition Theorem, 2. Multiplication Theorem, 3. Bernoulli's Theorem, 4. Bayes' Theorem

18.6.1 Addition Theorem

(A) Exclusive Events

If two events A and B are mutually exclusive and related to same group, the probability of the occurrence of either A or B is the sum of the individual probability of A and B. Thus,

p (AUB) = p (A) + p (B)

This theorem can be extended to three or more mutually exclusive events. Thus,

p (A or B or C) = p (A) + p (B) + p(C)

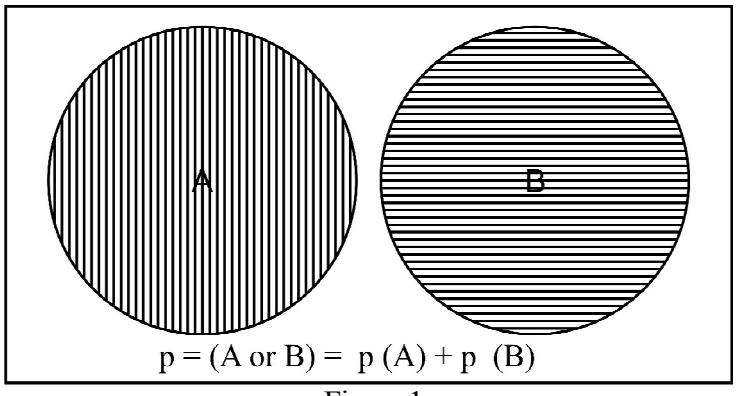


Figure 1

Illustration 6: A card is drawn out of a pack of 52 cards. Find the probability that it is a card of spade or of diamond.

Solution

p (A) = Probability of a card of spade = $\frac{13}{52}$

p (B) = Probability of a card of diamond = $\frac{13}{52}$

p (A or B) = Probability of a card of spade or diamond

= $\frac{13}{52} + \frac{13}{52} = \frac{26}{52}$ or $\frac{1}{2}$

Illustration 7: A bag contains 6 red, 4 black, 3 white and 2 green balls. What is the probability of getting a white or green ball at random in a single draw of one ball?

Solution:

Total balls = 6 + 4 + 3 + 2 = 15

Probability of getting white ball = $\frac{3}{15}$

Probability of getting green ball = $\frac{2}{15}$

Total probability for getting a white ball or green ball = $\frac{3}{15} + \frac{2}{15} = \frac{5}{15}$ or $\frac{1}{3}$

(B) Not Mutually Exclusive Events: When two events are not mutually exclusive on the whole or there are some common cases also, then the probability of the common cases should be subtracted from the sum of their individual probabilities. It can be expressed by the following formula

p (A or B) = p (A) + p (B) - p (AB)

In other words, p (A U B) = p (A) + p (B) - p (A∩B)

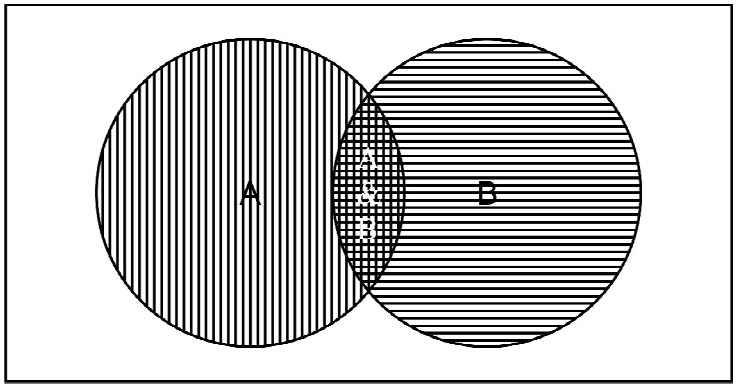


Figure 2

This theorem can be extended to more than two events also. Thus:

p (A or B or C) = p (A) + p (B) + p (C) - p (AB) - p (AC) - p (BC) + p (ABC)

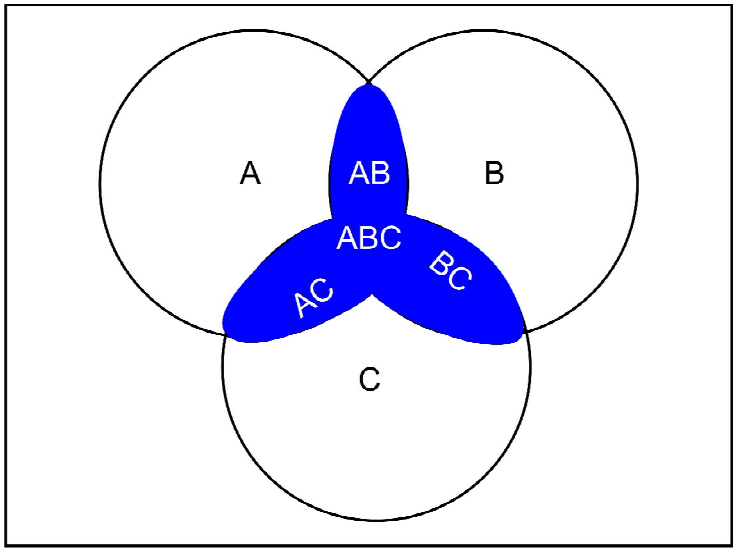


Figure 3

Illustration 8: A card is drawn out of a pack of 52 cards. What is the probability of drawing a card of Diamond or a Queen?

Solution:

Probability of drawing a card of diamond $= \frac{13}{52}$

Probability of drawing a queen $= \frac{4}{52}$

Probability of drawing the queen of diamond $= \frac{1}{52}$

$\therefore$ Probability that the card drawn is a card of diamond or a queen would be:

$$\frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} \text{ or } \frac{4}{13}$$

Illustration 9: Fifty balls marked serially from 1 to 50 were put into a bag. What is the probability that first drawn ball is of a marked number multiple to 5 or 8?

Solution:

Multiple of 5 = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 = 10

Multiple of 8 = 8, 16, 24, 32, 40, 48 = 6

But, one number i.e 40 is common, which should not be counted in double. Hence, the required probability would be $= \frac{10}{50} + \frac{6}{50} - \frac{1}{50} = \frac{15}{50} = \frac{3}{50}$

Illustration 10: In a randomly selected leap year, what is the probability that there are

(i) 53 Sundays, (ii) 53 Mondays

Solution:

We know that a leap year contains 366 days and there 52 full weeks and 2 days extra. These extra two days may be of different combinations:

(i) Sunday and Monday

(ii) Monday and Tuesday

(iii) Tuesday and Wednesday

(iv) Wednesday and Thursday

(v) Thursday and Friday

(vi) Friday and Saturday

(vii) Saturday and Sunday.

(i) Of these seven equally likely cases only 2 are such that contains Sunday. So probability of there being 53 Sundays is $= \frac{2}{7}$

(ii) Of these seven equally likely cases 2 combinations contain Monday's and 2 Combinations having Friday's. So the probability of either 53 Mondays or 53 Fridays is $= \frac{2}{7} + \frac{2}{7} = \frac{4}{7}$

18.6.2 Multiplication Theorem

(i) When events are independent:-

If two events A and B are independent, the probability that they both will occur is equal to the product of their individual probabilities. This is called Multiplication theorem of probability. Thus,

$$p(A \text{ and } B) \text{ or } p(AB) = p(A) \cdot p(B)$$

This rule can be extended to more than two events like;

$$p(A,B,C \text{ and.....}) = p(A) \cdot p(B) \cdot p(C) .....$$

Hint- When it is asked that both the events must occur together, then p(A and B) is to be calculated. In other words, multiplication rule is used. On the other hand when it is asked that either of the two events must happen then p(A and B) is to be calculated or additive model is used. If it is asked that certain event must happen simultaneously and there are certain events either of which can happen, both addition and multiplication theorems are required to be applied.

Illustration 11: An unbiased coin is tossed twice. Find out the probability of first getting the tail and next getting the head.

Solution

Probability for getting tail at first time $= \frac{1}{2}$

Probability for getting head at second time $= \frac{1}{2}$

$\therefore$ Combined probability $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

Probability of at least one event:

If we are given many independent events and out of which, the probability of happening of at least one event can be determined as follows,

p(occurrence of at least one event) = 1 - p (occurrence of none of the events)

In simple words, we can say that first of all we will find the combined probability of not happening of all these events and then this combined probability is deducted from 1.

Illustration 12: A problem in statistics is given to three students R, A and M, whose chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{2}{5}$ respectively. What is the probability that the problem will be solved?

Solution

(i) Probability that R solves the problem $= \frac{1}{2}$

$\therefore$ Probability that R does not solve the problem $= 1 - \frac{1}{2} = \frac{1}{2}$

Similarly, probability of A does not solve the problem $= 1 - \frac{1}{3} = \frac{2}{3}$

and probability of M not solving the problem $= 1 - \frac{2}{5} = \frac{3}{5}$

$\therefore$ Probability that none of the three is able to solve the problem is $= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{5} = \frac{1}{5}$

Probability that the problem is solved by at least one of them $= 1 - \frac{1}{5} = \frac{4}{5}$ or 80%

Illustration 13: From 6 Indians and 5 Japanese, a committee of 4 is to be formed. In how many ways can this be done when the committee consists of:

(i) All Indians, (ii) At least 2 Japanese

Solution

(i) When all the persons in the committee are Indians, we can select the committee members in the following ways:-

$= {}^6C_4 = \frac{\lfloor 6}{\lfloor 6-4 \rfloor 4} = \frac{\lfloor 6}{\lfloor 2 \rfloor 4}$ or $= \frac{6.5}{2.1} =$ 15 ways

(ii) Committee of 4 should include at least 2 Japanese, and it can be formed in the following ways:

(a) 2 Indians & 2 Japanese or ${}^6C_2 \cdot {}^5C_2$

(b) 1 Indian & 3 Japanese or ${}^6C_1 \cdot {}^5C_3$

(c) 4 Japanese or ${}^5C_4$

$\therefore$ Committee can be formed as:

$= ({}^6C_2 \cdot {}^5C_2) + ({}^6C_1 \cdot {}^5C_3) + ({}^5C_4)$

$= \left(\frac{6.5}{2.1} \cdot \frac{5.4}{2.1}\right) + \left(\frac{6}{1} \cdot \frac{5.4}{2.1}\right) + \left(\frac{5}{1}\right)$

$= (15 \cdot 10) + (6 \cdot 10) + (5) = 150 + 60 + 5$

$= 215$

(ii) When events are dependent -

Probability of these events is also called conditional probability. If A and B are two dependent events, the probability of their joint happening is p(A and B) = p(A) . p(B/A)

If B had already happened then,

p(A and B) = p(B) . p(A/B)

Illustration 14: Three cards are drawn consecutively from a pack of 52 cards one by one without replacement. Find the probability of drawing a jack, a queen and a king respectively.

Solution

Probability of drawing a jack $= \frac{4}{52}$

Probability of drawing a queen when first card is not replaced $= \frac{4}{51}$

Probability of drawing a king when both first and second cards are not replaced $= \frac{4}{50}$

So, required probability $= \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50}$

$= \frac{64}{132600} = \frac{8}{16575}$

18.6.3 Bernoulli's Theorems

If in trials of an experiment:-

(i) Success is given p and failure is denoted by q or (1-p),

(ii) Outcomes of every trial are independent,

(iii) Probability of success or failure remains the same in every trial, the probability of exactly 'r' successes in 'n' trials are determined by Bernoulli's theorem, which was propounded by Jacob Bernoulli.

Formula:

$$p(r) = {}^{n}C_{r} \cdot P^{r} q^{n-r}$$

Here, n = number of trials

r = number of success in the trials
p = probability of success in a trial and
q = 1 - p or the probability of failure in the trials

Illustration 15: What is the probability of getting exactly three heads when four coins are tossed?

Solution

Probability of getting heads $= \frac{1}{2}$

Probability of getting tails $= 1 - \frac{1}{2} = \frac{1}{2}$ No. of coins or n = 4

Probability of getting exactly three heads when four coins are tossed

$= {}^{n}C_{r} (p)^{r} \cdot (q)^{n-r}$ or ${}^{4}C_{3} (\frac{1}{2})^{3} \cdot (\frac{1}{2})^{4-3}$

$= {}^{4}C_{3} (\frac{1}{2})^{3} \cdot (\frac{1}{2}) = \frac{4!}{3!\,1!} (\frac{1}{2})^{3} \cdot (\frac{1}{2}) = 4 \cdot \frac{1}{16} = \frac{1}{4}$

18.6.4 Baye's Theorem

When on the basis of effects the causes are traced out, this is called inverse relationship. Probability in these types of relationship is based on Bayes' theorem. This concept is an extension of conditional probability. If on the basis of new information conditional probabilities are revised, this is done by using Bayes' theorem. Probabilities before revision by Bayes' rule are called as prior probabilities, because they are

determined before the information is taken into account. These prior probabilities are revised by using Bayes' theorem on the basis of sample information. These are called posterior probabilities. This concept is very useful in Business and Management to arrive at valid decisions in the face of uncertainties.

Formula -

$$P = \frac{p_m p_m}{p_1 p_1 + p_2 p_2 + p_3 p_3 \ldots\ldots\ldots\ldots + p_n p_n}$$

Illustration: 16

One bag contains 4 white balls and 4 black balls, second bag contains 3 white and 5 black balls and the third contains 4 white and 6 black balls. One black ball was drawn, what is the probability that it was drawn from the first bag?

Solution

$P(A_1)$ = Probability of drawing a black ball from first bag $= \frac{4}{8}$

$p(A_2)$ = Probability of drawing a black ball from second bag $= \frac{5}{8}$

$p(A_3)$ = Probability of drawing a black ball from third bag $= \frac{6}{10}$

p(B) Probability of selecting a bag $= \frac{1}{3}$

$$p(B \text{ and } A_1) = \frac{1}{3} \cdot \frac{4}{8} = \frac{4}{24}$$

$$p(B \text{ and } A_2) = \frac{1}{3} \cdot \frac{5}{8} = \frac{5}{24}$$

$$p(B \text{ and } A_3) = \frac{1}{3} \cdot \frac{6}{10} = \frac{6}{30}$$

$$p(B \text{ and } A_1) = \frac{p(B \text{ and } A_1)}{p(B \text{ and } A_1) + p(B \text{ and } A_2) + p(B \text{ and } A_3)}$$

$$= \frac{\frac{4}{24}}{\frac{4}{24} + \frac{5}{24} + \frac{6}{30}} = \frac{120}{414}$$

18.7 Mathematical Expectation

If the probability of happening an event is multiplied by expected numbers of outcomes or by expected value(getting money on happening an event), this is known as Mathematical Expectation.

Formula, E = PM

Here, E = Mathematical Expectation

P = Probability of an event

M = Money to be received on happening of an event.

If the events are more than one the formula is extended. Thus,

$$E = P_1 M_1 + P_2 M_2 + P_3 M_3 + \ldots\ldots\ldots\ldots P_n M_n$$

Illustration: 17

P and R enter into a bet according to which P will get Rs.400 if it rains on that day and will lose Rs.200 if it doesn't rain. The probability of raining on that day is 0.6. What is the mathematical expectation of P.

Solution

Probability of raining = 0.6

Probability of not raining = 0.4

Expected value if it rains = 0.6 . 400= Rs 240

Expected value if it does not rain = 0.4 . 200= Rs 80

$\therefore$ Mathematical Expectation = Rs 240 - Rs 80 = Rs 160

18.8 Importance of Probability Theory

Though probability theory was developed to solve the problems of gamblers but now it is useful not only in solving mathematical problems but also social, business, economical, and political problems. If we have to estimate the future or take some decisions in uncertainty, probability theory provides all the answers. In statistics law of inertia of large numbers and law of statistical regularity are based on it. Knowledge of probability is a must for using statistical decision theory, theoretical frequency distribution, statistical inferences, test of hypothesis and test of significance. Based on the future expectation or probability, important decisions related with sales, production, finance etc. are taken. In insurance business calculation of loss or risk of life is done by expectations to calculate premium. Infact, probability has become a part of everyone's life to make better decisions.

18.9 Summary

The probability of a given event is an expression of likelihood or chance of happening of an event. It is a numerical measure with a value between 0 and 1. This theory provides us a mechanism for measuring and analysing uncertainties associated with events. Probability models can be very useful for making predictions.

Probability can be objective or subjective. If two events A and B are mutually exclusive, the probability of the occurrence of either A or B is calculated by Addition theorem, if events are not mutually exclusive then addition rule must be modified. If events A and B are independent events, then the probability that they both will occur is calculated by multiplication theorem. If events A and B are so related that occurrence of B is affected by the occurrence of A, then this is conditional probability. This concept can be extended to revise probabilities. In other words, when on the basis of effects the causes are traced out, it is termed as inverse relationship. The procedure of revising is termed as Bayes' theorem. This theorem is useful in order to improve the prior probabilities on the basis of new information.

Probability has become an indispensable tool for all types of formal studies that involve uncertainty. It is used for solving many problems of everyday life, scientific investigations, managerial decisions on planning and control econometric models are based on probability theorem. Even before learning statistical decision procedure knowledge of probability theory is essential.

18.10 Key Words

Probability: A numerical measure of the likelihood that a particular event will occur.

Event: An event is an outcome or set of outcomes of an activity.

Mutually Exclusive Events: If both events cannot occur at the same time as outcome of a single experiment, these events are mutually exclusive.

Independent and Dependent Events: When the outcomes of one does not affect and is not affected by

the other, these two events are called independent events. Dependent events are those in which occurrence of one event affects the probability of occurrence of other event.

Equally Likely Events: If an event occurs the same number of times as other events occur, they are equally likely events.

Addition Theorem: If two events A and B are mutually exclusive the probability of the occurrence of either A or B is the sum of the individual probability of A and B.

Multiplication theorem: If two events A and B are independent the probability that they both will occur is equal to the product of their individual probability.

18.11 Self Assessment Questions

1. Define probability and explain importance of this concept.
2. Briefly explain the different approaches of probability and give examples illustrating the application of these theorems.
3. Explain the terms mutually exclusive events and independent events.
4. State the addition and multiplication theorems of probability and give examples illustrating the application of these theorems.
5. State and explain Bayes' theorem and bring out its importance in probability theory.
6. In how many ways can two prizes be awarded to five contestant if both the prizes-

   (a) May not be given to the same person and

   (b) May be given to the same person.
7. How many arrangements can be formed with letters of the words (i) GANGAPUR (ii) STATISTICS.

   (i) 10080 (ii) 50400
8. Two dices are tossed. What is the probability that sum shown will be 7 or 11?

   (2/9)
9. (a) Three perfect coins are tossed together. What is the probability of getting at least one head?

   (b) From a bag containing 10 black and 20 white balls, a ball is drawn at random. What is the probability that it is black.

   (a) 7/8, (b) 1/3
10. A problem in accounts is given to five students A,B,C,D and E. Their chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$ and $\frac{1}{6}$. What is the probability that the problem will be solved?

    (5/6)
11. One bag contains 4 red and 2 green balls. Another contains 3 red and 5 green balls. If one ball is drawn from each bag, find the probability that (a)both are red, (b)both are green and (c) one is red and one is green?

    (a) 1/4 (b) 5/24, (c) 13/24
12. Three groups of workers contain 3 men and one women, 2 men and 2 women, and 1 man and 3

women respectively. One worker is selected at random from each group. What is the probability that the group selected consists of 1 man and 2 women?

(13/32)

13. A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black.

(3/28)

14. A factory has two machines. Past records show that machine 1 produces 30% of the items of output and machine 2 produces 70% of the items. Further 5% of the items produced by the machine 1 were defective and only 1% produced by machine 2 were defective. If a defective item is drawn at random, what is the probability that the defective item was produced by the machine 1 or machine 2?

(Machine 1- 68.2%
Machine 2 - 31.8%)

15. There are three urns. Urn A contain 8 red and 7 green marbles. Urn B has 5 red and 8 green marbles. One red marbles is drawn from one of the urn. Find out the probability that it came from (a) Urn A (b) Urn C?

(Urn A -104/309,
Urn C - 130/309)

18.12 Reference Books

1. Gupta, S.P.. Statistics, Sultan Chand & Sons.
2. Sharma, J K, Business Statistics, Pearson Education.
3. Chandan, J S, Business Statistics, Vikas Publishing House Pvt Ltd, New Delhi.
4. Hooda R P, Statistics for Business and Economic, Macmillan, New Delhi.
5. Levin and Rubin. Statistics for management, Prentice Hall of India Ltd., New Delhi.

BBA-05

Vardhaman Mahaveer Open University, Kota

Fundamentals of Business Statistics

Course Development Committee

Chairman
Prof. (Dr.) Naresh Dadhich
Vice-Chancellor
Vardhaman Mahaveer Open University, Kota

Convener and Members

Subject Convener
Dr. Anurodh Godha
Assistant Professor, Department of Commerce,
Vardhaman Mahaveer Open University, Kota

Members:

1. **Prof. Parimal H. Vyas**
   Professor & Head,
   Deptt. of Commerce and Business Management,
   Faculty of Commerce,
   The M.S. University of Baroda, Vadodara (Gujarat)
2. **Prof. R.C.S. Rajpurohit**
   Professor & Head,
   Deptt. of Business Administration,
   J.N.V. University, Jodhpur (Raj.)
3. **Prof. N.D. Mathur**
   Professor & Head,
   Deptt. of Management Studies,
   Central University of Rajasthan,Kishangarh-Ajmer (Raj.)
4. **Prof. Rajeev Jain**
   Director & Dean,Faculty of Management Studies,
   J.R.N. Rajasthan Vidyapeeth University, Udaipur (Raj.)
5. **Prof. Shyam Gopal Sharma**
   Senior Most Professor & Former Head,
   Deptt. of ABST,
   University of Rajasthan, Jaipur (Raj.)
6. **Prof. M.C. Govil**
   Principal,
   Govt. Women Engineering college, Ajmer (Raj.)
7. **Prof. Navin Mathur**
   Professor, A.S. to Vice-Chancellor,
   Faculty of Management Studies,
   University of Rajasthan, Jaipur (Raj.)
8. **Prof. Karunesh Sexsena**
   Director & Chairman,
   Faculty of Management Studies,
   Mohanlal Sukhadia University, Udaipur (Raj.)

Editing and Course Writing

Editor:

Prof. S.C. Bardia
Professor, Department of ABST
University of Rajasthan, Jaipur (Rajasthan)

Writers:

Dr. Namita Jalan (Unit No. 1, 2)
Lecturer in ABST,
Vardhaman Kanya Mahavidhyalaya, Beawar (Rajasthan)

Dr. Suraksha Sharma (Unit No. 3)
Lecturer in ABST,
Vedic Kanya P. G. Mahavidhyalaya, Jaipur (Rajasthan)

Dr. Manish Jain (Unit No. 4)
Principal,
Siddharth Institute of Modern Management, Jaipur (Raj.)

Dr. Seema Agrawal (Unit No. 5)
Head, Department of ABST,
Kanoria P. G. Mahila Mahavidhyalaya, Jaipur (Rajasthan)

Dr. Shiv Prasad (Unit No. 6, 7)
Associate Professor, Deptt. of Management Studies,
Maharshi Dayanand Saraswati University, Ajmer (Raj.)

Dr. Sonia Agarwal (Unit No. 8, 9)
Assistant Professor, Department of Management,
Banasthali University, Jaipur (Rajasthan)

Dr. Anurodh Godha (Unit No. 10, 17)
Assistant Professor, Department of Commerce,
Vardhaman Mahaveer Open University, Kota (Rajasthan)

Dr. Prerna Jain (Unit No. 11, 12, 14, 18)
Lecturer in ABST,
Govt. College, Ajmer (Rajasthan)

Dr. Deepika Upadhyaya (Unit No. 13)
Assistant Professor, Department of Management Studies,
Maharshi Dayanand Saraswati University, Ajmer (Raj.)

Dr. Meenu Maheshwari (Unit No. 15, 16)
AICWA & Head, Deptt. of Commerce and Management,
University of Kota, Kota (Rajasthan)

Academic and Administrative Management

Prof. (Dr.) Naresh Dadhich	**Prof. M.K. Ghadoliya**	**Mr. Yogendra Goyal**
Vice-Chancellor	Director (Academic)	In charge
Vardhaman Mahaveer Open University, Kota	Vardhaman Mahaveer Open University, Kota	Material Production and Distribution Department

Course Material Production

Mr. Yogendra Goyal
Assistant Production Officer
Vardhaman Mahaveer Open University, Kota

PRODUCTION Feb. 2011 **ISBN -13/978-81-8496-272-7**

Printed and published on behalf of V.M. Open University, Kota by Registrar
Printers : The Pooja, Kota/Feb 2011/500

Vardhaman Mahaveer Open University, Kota

CONTENTS

Fundamentals of Business Statistics