Centre for Distance & Online Education
(CDOE)

BACHELOR OF COMMERCE

BCOM 302

BUSINESS STATISTICS-I

Guru Jambheshwar University of Science & Technology, Hisar – 125001

Contents

Subject : **Business Statistics-1**	
Course Code : **BCOM 302**	Author : **Dr. Pradeep Gupta**
Lesson No. : 1	Vetter: **Prof. B. S. Bodla**

Statistics: Scope, Usefulness and Limitation

Structure

1.0 Learning Objectives

After going through this lesson, you will be able to

- Explain the concept of statistics

- find the scope of statistics
- find the usefulness of statistics
- find the limitations of statistics
- explain the distrust of statistics

1.1 Introduction

Life in the modern world is inexuricably bound with the notions of number, counting and measurement. One day try to think of a community that cannot count or take measurements and yet is concerned with such acts as selling and buying, carrying on bank transactions, operating locomotives, cars, ships, aircraft and taking part in government. The overriding importance of numerical data in modern life will then be all too apparent. Statistics is being used both is a singular noun and a plural noun.

Statistics, as a plural noun, is used to mean numerical data which arise from a host of uncontrolled, and mostly unknown, causes acting together. It is in this sense that the term statistics is used when our daily newspapers give vital statistics, crime statistics or soccer statistics of Calcutta, or when the Food Minister in the Lok Sabha quotes statistics of sugar exports or those of food grain production.

Used as singular, statistics is a name for the body of scientific methods which are meant for the collection, classification, tabulation, analysis and interpretation of numerical data. But modern literature on the subject does away with any such distinction.

1.2 Concept of Statistics

Statistics is not a new discipline but as old as the human society itself. In the old days statistics was regarded as the ’Science of Statecraft’ and was the by-product of the administrative activity of the State. It has been the traditional function of the governments to keep records of population, births, deaths, taxes crop yields and many other types of activities. Counting and measuring these events may generate much kind of numerical data.

The word ’statistics’ comes from the Italian word ’statista’ (meaning "Statesman") or the German word ’Statistik’ each of which means a Political State. It was first used by Professor Gottfried in 1749 to refer

to the subject matter as a whole. The science of statistics is said to have originated from two main sources.

(a) Government Record: This is the earliest foundation because all cultures with a recorded history had recorded statistics, and the recording, as far as is known, was done by agents of the government for governmental purpose. Since statistical data were collected for governmental purpose, statistics was then described as the 'science of kings' or 'the science of statecraft'.

(b) Mathematics: Statistics is said to be a branch of applied mathematics. The present body of statistical methods, particularly those concerned with drawing inferences about population from a sample is based on the mathematical theory of probability.

The following are the two main factors which are responsible for the development or statistics in modern time:

(a) Increased demand for statistics: In the present century considerable development has taken place in the field of business and commerce, governmental activities and science. Statistics help in formulating suitable policies, and as such its need is increasingly felt in all these spheres.

(b) Reduced cost of statistics: The time and cost of collecting data are very important limiting factors in the use of statistics. However, with the development of electronic machines, such as calculators, computers etc. the cost of analyzing data has considerably gone down. This has led to the increasing use of statistics in solving various problems. Moreover, with the development of statistical theory the cost of collecting and processing data has gone. For example, considerable advance has been made in the sampling techniques which enable us to know the characteristics of the population by studying only a part of it.

1.2.1 Definition of Statistics

The purpose of definition is to lay down precisely the meaning, the scope and the limitations of a subject. There are many definitions of the term 'statistics'. A few definitions are analytically examined below:

(1) Webster defined statistics as "the classified facts representing the conditions 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".

(2) Yule and Kendall defined statistics as "By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes".

(3) Croxton and Cowden have given a very simple and concise definition of statistics. In their view "Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data".

(4) According to Berenson and Levin, "The science of statistics can be viewed as the application of the scientific method in the analysis of numerical data for the purpose of making rational decisions".

(5) Boddington defines statistics as "the science of estimates and probabilities".

(6) According to Lincon L. Chao, "Modern statistics refers to a body of methods and principles that have been developed to handle the collection, description, summarisation and analysis of numerical data. Its primary objective is to assist the researcher in making decisions or generalizations about the nature and characteristics of all the potential observations under consideration of which the collected data form only a small part".

All the above definitions are less comprehensive than the one given by Prof. Horace who defined statistics as follows:

"By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other".

(i) Statistics are aggregate of facts: Single and isolated figures are not statistics for the simple reason that such figures are unrelated and cannot be compared. To illustrate, if it is stated that the income or Mr. A is Rs. 1, 00,000 per annum, this would not constitute statistics although it is numerical state of fact. Similarly, a single figure relating to production, sale, birth, employment, purchases, accident etc. cannot be regarded statistics although aggregates of such figures would be statistics because of their comparability and relationship as part of common phenomenon.

(ii) Statistics are affected to a marked extent by multiplicity of causes: Facts and figures are affected to a considerable extent by a number of forces operating together. For example,

statistics of production of rice are affected by the rainfall, quality of soil, seeds, manure, method of cultivation etc.

(iii) Statistics are numerically expressed: All statistics are numerical statements of facts i.e. expressed in numbers. Qualitative statements such as 'the population of India is rapidly increasing', or 'the production of wheat is not sufficient' do not constitute statistics. The reason is that such statements are vague and one cannot make at anything from them. On the other hand, the statement 'The estimated population of India at the end of VIIth plan is 803 million' is a statistical statement.

(iv) Statistics are enumerated or estimated according to a reasonable standard of accuracy: Facts and figures about any phenomenon can be derived in two ways, viz by actual counting and measurement or by estimate. Estimates cannot be as precise and accurate as actual counts or measurements. The degree of accuracy desired largely depends or measurements. The degree of accuracy desired largely depends upon the nature and object of the enquiry. For example, in measuring heights of persons even 1/10th of a cm is material whereas in measuring distance between two places, say Madras and Calcutta, even fraction of a kilometer can be ignored. However, it is important that reasonable standards of accuracy should be attained; otherwise numbers may be altogether misleading.

(v) Statistics are collected in a systematic manner: B efore collecting statistics a suitable plan of data collecting should be prepared and the work carried out in a systematic manner. Data collected in a haphazard manner would very likely lead to fallacious decisions.

(vi) Statistics are collected for a pre-determined purpose: The purpose of collecting data must be decided in advance. The purpose should be specific and will define. A general statement of purpose is not enough. For example, if the objective is to collect data on prices, it would not serve any useful purpose unless one knows whether he wants to collect data on wholesale or retail prices and what are the relevant commodities in view.

(vii) Statistics should be placed in relation to each other: If numerical facts are to be called statistics, they should be comparable. Statistics data are often compared period-wise or region-wise. For example, the per capita income of India at a particular point of time may be compared with that of earlier years or with the per capita income of other countries, say U.S.A., UK, China

etc. Valid comparisons can be made only if the data are homogeneous i.e. relate to the same phenomenon or subject and only likes are compared with likes. It would be meaningless to compare the height of elephants with the height of human beings.

In the absence of the above characteristics, numerical data cannot be called statistics.

1.2.2 Scope of Statistics

(i) Statistics bring definiteness and precision in conclusions by expressing them numerically. It is the quality of definiteness which is responsible for the growing universal applications of statistical methods. The conclusions stated numerically are definite and hence more convincing than conclusions stated qualitatively. This fact can be readily understood by a simple example. In an advertisement, statements expressed numerically have greater attention and more appealing than those expressed in a qualitative manner. The caption 'we have sold more. T.Vs this year', is certainly less attractive than 'Record Sale of 15,000 T.V. in 1998 as compared to 10,000 in 1997'. The latter statement emphasizes in a much better manner the growing popularity of the advertises T.Vs.

(ii) Statistics make data comprehensible to the human mind by simplifying and summarizing it. Statistics simplifies unwieldy and complex mass of data and presents them in such a manner that they at once become intelligible. The complex data may be reduced to totals, averages, percentage etc. and presented either graphically or diagrammatically. This derives help to understand quickly the significant characteristics of the numerical data, and consequently save from a lot of mental strain. Single figures in the form of averages and percentages can be grasped more easily than a mass of statistical data comprising thousands of facts. Similarly, diagrams and graphs, because of their greater appeal to the eye and imagination tender valuable assistance in the proper understanding of numerical data. Time and energy of business executives are thus economized, if the statistician supplies them with the results of production, sale and finances in a condensed form.

(iii) Statistics facilitate comparisons in the data. Certain facts, by themselves, may be meaningless unless they are capable of being compared with similar facts at other places or at other periods in times. For example, we estimate the national income of India not essentially for

the value of that fact itself, but mainly in order that we may compare the income of today with that of the past and thus draw conclusions as to whether the standard of living of the people is on the increase, decrease or is stationary. It is with the help of statistics that the cost accountant is able to compare the actual accomplishment (in terms of cost). Some of the modes of comparison provided by statistics are: Totals, ratios, averages or measure of central tendencies, graphs & diagrams and coefficients. Statistics thus 'serves as a scale in which facts in various combinations are weighed and valued'.

(iv) Statistics studies and establishes among the variables. Certain statistical measures such as coefficient of correlation, regression etc. establishes relationship between different types of data. For example, it is possible to observe the relationship between income and expenditure, export and forex reserves etc.

(v) Statistics helps in formulating and testing hypothesis. Statistical methods are extremely useful in formulating and testing hypothesis and to develop new theories. For examples the hypothesis that a new drug is effective in checking malaria, will require the use of statistical technique of association of attributes.

(vi) Statistics helps in prediction. 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. The statistical techniques for extrapolation, time series etc. are highly useful for forecasting future events.

(vii) Statistics helps in the formulation of suitable policies. Statistics help in formulating policies in social, economic and business fields. Various government policies in the field of planning taxation, foreign trade, social security etc. are formulated on the basis of analysis of statistical data and the inferences drawn from them. For example, vital statistics comprising birth and morality rates help in assessing future growth in population. This information is necessary for designing any scheme of family planning. Similarly, the rate of dearness allowance to be given to the employees is calculated with the help of index numbers.

(viii) Statistics draws inferences for taking decisions. Statistical tests are devised to help in drawing valid inference in regard to the nature and characteristics of the universe on the basis of the study of the sample. It can also be the other way when the nature of the sample is judged

on the basis of the parameters based on the study of the universe. The validity of such inferences depends on the type of statistical methods employed for the purpose.

(ix) Statistics endeavors to interpret conditions. Statistics render useful service by enabling the interpretation of condition, by developing possible causes for the results described. For example, if the production manager discovers that a certain machine is turning out some articles which are not of standard specifications, he will be able to find statistically if this condition is due to some defects in the machine or whether such a condition is normal.

(x) Statistics measures uncertainty. Statistical methods help not only in ascertaining the chance of occurrence of an event but also in finding out the total effect of an uncertain event if the consequences of various occurrences are known. Both objective and subjective probability estimates are employed depending upon the nature of the enquiry.

(xi) Statistics enlarges individual experience. A proper function of statistics indeed is to enlarge individual experience. Many fields of knowledge would have remained closed to mankind, without the efficient and useful techniques of statistical analysis.

1.2.3 Usefulness of statistics

Statistical methods have become useful tools in the world of affairs. Economy and a high degree of flexibility are the important qualities of statistical methods that render them especially useful to businessmen and scientists.

Statistics and Business: Statistical information is needed from the time the business is launched till the time of its exit. At the time of the floatation of the concern facts are required for the purpose of drawing up the financial plan of the proposed unit. All the factors that are likely to affect judgment on these matters are quantitatively weighed and statistically analyzed before taking the decisions.

Statistical methods of analysis are helpful in the marketing function of an enterprise though enormous help in market research, advertisement campaigns and in comparing the sales performances. Statistics also directs attention towards the effective use of advertising funds.

Correlation and regression analysis help in the estimation of relationships between dependent and one or more independent variables e.g. relationships are established between market demand and per capita

income, inputs and outputs etc.

They theory and techniques of sampling can be used in connection with various business surveys with a considerable saving in time and money. Likewise these techniques are now being extensively used in checking of accounts.

Statistical quality control is now being used in industry for establishing quality standards for products, for maintaining the requisite quality, and for assuring that the individual lots sold are of a given standard of acceptance.

The use for statistical information in the smooth functioning of an undertaking increases along with its size. The bigger the concern the greater is the need for statistics.

Statistics is thus a useful tool in the hands of the management. But it must be remembered that no volume of statistics can replace the knowledge and experiences of the executives. Statistics supplements their knowledge with more precise facts than were hitherto available.

Statistics & Economics*:* Statistical data and methods of statistical analysis render valuable assistance in the proper understanding of the economic problems and the formulation of economic policy. Economic problems almost always involve facts that are capable of being expressed numerically, e.g. volume of trade, output of industries - manufacturing, mining and agriculture - wages, prices, bank deposits, clearing house returns etc. These numerical magnitudes are the outcome of a multiplicity of causes and are consequently subject to variations from time to time, or between places or among particular cases. Accordingly, the study of economic problem is specially suited to statistical treatment.

The development of economic theory has also been facilitated by the use of statistics. Statistics is now being used increasingly not only to develop new economic concepts but also to test the old ones. The increasing importance of statistics in the study of economic problem has resulted in a new branch of study called Econometrics.

Statistics and Biology*:* Statistics is being used more and more in biological sciences as an aid to the intelligent planning of experiments, and as a means of assuring the significance of the results of such experiments. Experiments about the growth of animals under different diets and environments, or the crop yields with different seeds, fertilizers and types of soil are

frequently designed and analyzed according to statistical principles.

Statistics and physical sciences: Statistics is not much in use in the fields of Astronomy, Geology and Physics. This is due mainly to their relatively high precision of measurements. Statistics has not made any progress in physical sciences beyond the calculation of standard error, and fittings of curves.

Statistics and computers: The development of statistics has been closely related to the evolution of electronic computing machinery. Statistics is a form of data processing, a way of converting data into information useful for decision making. A huge mass of raw data, of related and unrelated nature, derived from internal and external sources of different period of time can be organized and processed into information by computers with accuracy and high speed. The computers can make complex computations, analysis, comparisons and summarizations. Though humans can do the processing, the computer's ability to process huge data is phenomenal, considering its speed, reliability and faithfulness in perfectly following the set of instructions.

The input data in the computer can be processed into a number of different outputs and for a variety of purposes. The system is so organized that managers at different levels and in different activity units are in a position to obtain information in whatever form they want, provided that relevant 'programmes' or instructions have been designed for the purpose. However, the output from a computer is only as good as the data input. 'Garbage In Garbage Out' is an adage familiar to computer users. This warning applies equally to statistical analysis. Statistical decisions based on data are no better than the data used.

As statisticians devise new ways of describing and using data for decisions, computer scientists respond with newer and more efficient ways of performing these operations. Conversely, with the evolution of more powerful computing techniques, people in statistics are encouraged to explore new and more sophisticated methods of statistical analysis.

Statistical Analysis Packages

Statistical Analysis Packages are preprogrammed with all the specialized formulas and built-in procedures a user may need to carry out a range of statistical studies. Statistical programs can:

- Accept data from other sources.

- Add or remove data items, columns or rows.
- Sort, merge and manipulate facts in numerous ways.
- Perform analysis on single and multiple sets of data.
- Convert numeric data into charts and graphs that people can use to grasp relationships, spot patterns and make more informed decisions.
- Print summary values and analysis results.

For a period of at least twenty years, groups of standardized statistical programs assembled as a collection or "package" have been available from various software developers. Recently, there has been a widespread development of statistical packages for use on a microcomputer. Certain packages that were previously available only for mainframe and minicomputers, (such as SAS, SPSS and Minitab) are now available in microcomputer versions, and many new packages (such as STATGRAPHICS, SYSTAT, MYSTAT) have been specifically developed for microcomputer use. The easy and relatively inexpensive access to this type of software has led to its ever-increasing use for business applications.

1.2.4 Limitations of statistics

Though the science of statistics has been profitably applied to an increasingly large number of problems, it has its own limitations and is at times misused by interested people who restrict its scope and utility. According to Newsholme, "It (Statistics) must be regarded as an instrument of research of great value, but having severe limitations which are not possible to overcome and as such they need our careful attention."

The following are some of the important limitations of statistics.

(i)Statistics does not study qualitative phenomenon: Statistics deals with only those subject of inquiry which are capable of being quantitatively measured and numerically expressed. This is an essential condition for the application of statistical methods. Now all subjects cannot be expressed in numbers. Health, poverty, intelligence (to name only a few) is instances of the objects that defy the measuring rod, and hence are not suitable for statistical analysis. The efforts are being made to accord statistical treatment to subjects of this nature also. Health of the people is judged by a study of the death rate, longevity of life and prevalence of any disease or diseases. Similarly intelligence of the students may be compared on the basis of the marks obtained by them in a class test. But these are only indirect

methods of approaching the problem and subsidiary to quite a number of other considerations which cannot be statistically dealth with.

(ii) Statistics does not study individuals: Statistics deals only with aggregates of facts and no importance is attached to individual items. Individual items, taken separately, do not constitute statistical data and are meaningless for any statistical inquiry. For example, the individual figures of agriculture production, industrial output or national income of any country for a particular year are meaningless, unless these figures enable comparison with similar figures for other countries and in the same country these are given for a number of years.

(iii) Statistical data is only approximately and not mathematically correct: Greater and greater emphasis is being laid on sampling technique of collecting data. This means that by observing only a limited number of item we make an estimate of the characteristics of the entire population. This system works well so long as the mathematical accuracy is not essential. But when exactness is essential statistics will fail to do the job.

(iv) Statistics is only one of the methods of studying a problem: Statistical tools do not provide the best solution under all circumstances. Very often, it is necessary to consider a problem in the light of a country's culture, religion and philosophy, Statistics cannot be of much help in studying such problems. Hence statistical conclusions must be supplemented by other evidences.

(v) Statistics can be misused: The greatest limitation of statistics is that it is liable to be misused. The misuse of statistics may arise because of several reasons. For example, if statistical conclusions are based on incomplete information, one may arrive at fallacious conclusions. Thus the argument that drinking beer is bad for longevity because 99% of the persons who take beer die before the age of 100 years is statistically defective, since we were not told what percentage of persons who do not drink beer die before reaching that age. Statistics are like clay and they can be moulded in any manner so as to establish right or wrong conclusions.

1.3 Distrust of statistics

It is a general belief that "statistics can prove anything." This statement is partly true and false. It is false because mere statistics should not be taken for granted without proper verification. It is true because statistics is often used by unscrupulous people to achieve their personal ends. This results in loss of faith or confidence on statistics or in causing distrust of statistics.

Distrust of statistics literally means lack of trust in statistical data, statistical analysis and the conclusions derived from it. The following reasons account for such views about statistics.

- Facts based on figures are more convincing. But these figures can be manipulated according to one's wishes. This misguides public causing distrust in statistics.
- They can be manipulated in such a manner as to establish foregone conclusions.
- The wrong representation of even correct figures can mislead a reader. Sometimes statistical analyses are misinterpreted causing distrust in statistics. Supposing the mortality rates of patients are more in Indian hospitals. From this one way wrongly conclude that it is safer to treat the patients at home. This type of misinterpretation also causes distrust in statistics.

Statistics are useful tools. One uses them according to his knowledge and experience. Use of statistics makes a statement more convincing. But its misuse causes distrust. So it is necessary that people should be adequately prepared to know the reality or to shift the truth from untruth, good statistics from bad statistics. Thus while working with statistics one should not only avoid outright falsehoods but be alert to detect possible distortion of the truth.

1.4 Check Your Progress

There are some activities to check your progress. Answer the followings:

1. Statistics are affected to a marked extent by multiplicity of ……
2. Statistics helps in the …….. of suitable policies.
3. Statistics bring definiteness and ……in conclusions by expressing them numerically.
4. Distrust of statistics literally means lack of trust in statistical data, statistical analysis and

the …… derived from it.

5. Statistical data is only…… and not mathematically correct.

1.5 Summary

Statistics is being used both is a singular noun and a plural noun. Statistics, as a plural noun, is used to mean numerical data which arise from a host of uncontrolled, and mostly unknown, causes acting together. Used as singular, statistics is a name for the body of scientific methods which are meant for the collection, classification, tabulation, analysis and interpretation of numerical data. "By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other". Statistics bring definiteness and precision in conclusions by expressing them numerically. Statistics make data comprehensible to the human mind by simplifying and summarizing it. Statistics facilitate comparisons in the data Statistics studies and establishes among the variables. Statistics helps in formulating, testing hypothesis, prediction and formulation of suitable policies. It is used in every field like business, economics, biology, physical science and computer etc. But there are some limitations of it like Statistics does not study qualitative phenomenon and on the individual basis. It is only approximately and not mathematically correct. It can also be misused Statistics are useful tools. One uses them according to his knowledge and experience. Use of statistics makes a statement more convincing. But its misuse causes distrust. So it is necessary that people should be adequately prepared to know the reality or to shift the truth from untruth, good statistics from bad statistics.

1.6 Keywords

Statistics: It mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.

Distrust of Statistics: It means lack of trust in statistical data, statistical analysis and the conclusions derived from it.

1.7 Self-Assessment Test

Q1. Define statistics. Also discuss the applications of statistics in business decision making.

Q2. Discuss the functions and limitations of statistics.

Q3. "Statistical methods are most dangerous tools in the hands of the expert" Elucidate.

Q4. "Statistics are numerical statement of facts but all facts numerically stated are not statistics" Comment upon the statement and state briefly which numerical statements of facts are not statistics.

Q5. How the computers can be helpful in making statistical decision?

1.8 Answer to Check Your Progress

1. Causes

2. Formulation

3. Precision

4. Conclusions

5. Approximately

1.9 References/Suggested Readings

Gupta, S. P.: Statistical Methods, Sultan chand and sons, New Delhi.

Kumar, S.: Practical Statistics, Sultan chand and sons, New Delhi

Levin, R. and David, S. R.: Statistics for Management, Prentice Hall, New Delhi.

Gupta, C. B.: Introduction to Statistical Methods, Ram Prashad, New Delhi.

Sancheti, D. C. and Kapoor, V. K.: Business Statistics.

Agarwal, B. L.: Basic Statistics, New age International.

Kapur, S. K.: Elements of Practical Statistic, Oxford & IBH Publishers.

Subject : **Business Statistics-1**	
Course Code : **BCOM 302**	Author : **Ms. Poonam**
Lesson No. : 2	Vetter: **Prof. Suresh K. Mittal**

COLLECTION OF DATA

Structure:

2.0 Learning Objectives

After going through this lesson, you will be able to:

- Know the different methods of data collection

- Understand the methodology of collecting primary data
- Define a questionnaire and its characteristics
- Understand the steps involved in questionnaire designing
- Know designing survey research
- Understand the methodology of collecting secondary data

2.1 Introduction

The facts and figures which can be numerically measured are studied in statistics. Numerical measures of same characteristic are known as observation and collection of observations is termed as data. Data are collected by individual research workers or by organization through sample surveys or experiments, keeping in view the objectives of the study. The data collected may be: Primary Data, Secondary Data. The difference between primary and secondary data in Statistics is that Primary data is collected first hand by a researcher (organization, person, authority, agency or party etc) through experiments, surveys, questionnaires, focus groups, conducting interviews and taking (required) measurements, while the secondary data is readily available (collected by someone else) and is available to the public domain through publications, journals and newspapers.

2.2 Primary Data Collection Methods

Many times due to inadequacy of data or stale information, the need arises for collecting a fresh firsthand information. In marketing research, there are broadly two ways by which primary information can be gathered namely, observation and communication.

Benefits of Primary Data

Benefits of Primary data cannot be neglected. A research can be conducted without secondary data but a research based on only secondary data is least reliable and may have biases because secondary data has already been manipulated by human beings. In statistical surveys it is necessary to get information from primary sources and work on primary data: for example, the statistical records of female population in a country cannot be based on newspaper, magazine and other printed sources. One such source is old and secondly they contain limited information as well as they can be misleading and biased.

Validity: Validity is one of the major concerns in a research. Validity is the quality of a research that

makes it trustworthy and scientific. Validity is the use of scientific methods in research which make it logical and acceptable. Using primary data in research can improves the validity of research. Firsthand information obtained from a sample that is representative of the target population will yield data that will be valid for the entire target population.

Authenticity: Authenticity is the genuineness of the research. Authenticity can be at stake if the researcher invests personal biases or uses misleading information in the research. Primary research tools and data can become more authentic if the methods chosen to analyze and interpret data are valid and reasonably suitable for the data type. Primary sources are more authentic because the facts have not been overdone. Primary source can be less authentic if the source hides information or alters facts due to some personal reasons. There are methods that can be employed to ensure factual yielding of data from the source.

Reliability: Reliability is the certainty that the research is enough true to be trusted on. For example, if a research study concludes that junk food consumption does not increase the risk of cancer and heart diseases. This conclusion should have to be drawn from a sample whose size, sampling technique and variability is not questionable. Reliability improves with using primary data. In the similar research mentioned above if the researcher uses experimental method and questionnaires the results will be highly reliable. On the other hand, if he relies on the data available in books and on internet he will collect information that does not represent the real facts.

Limitations of Primary Data Collection

One limitation of primary data collection is that it consumes a lot of time. The researchers will need to make certain preparations in order to handle the different demands of the processes and at the same time, manage time effectively. Besides time consumption, the researchers will collect large volumes of data when they collect primary data. Since they will interact with different people, they will end up with large volumes of data, which they will need to go through when analyzing and evaluating their findings. The primary data also require the greater proportion of workforce to be engaged in the collection of information and analysis, which enhances complexity of operations. There is requirement of large amount of resources to collect primary data. There are several methods of collecting the primary data, which are as follows:

- Observation Method

- Interview Method
- Through Questionnaires
- Through Schedules

Other methods such as warranty cards, distributor audits, pantry audits, consumer panels, using mechanical devices, through projective techniques, deep interviews and content analysis.

2.2.1 Observation Method

In the observation method, only present/current behavior can be studied. Therefore, many researchers feel that this is a great disadvantage. A causal observation could enlighten the researcher to identify the problem. Such as the length of the queue in front of a food chain, price and advertising activity of the competitor etc. Observation is the least expensive mode of data collection.

E*xample:* Suppose a Road Safety Week is observed in a city and the public is made aware of advance precautions while walking on the road. After one week an observer can stand at a street corner and observe the number of people walking on the footpath and those walking on the road during a given period of time. This will tell him whether the campaign on safety is successful or unsuccessful. Sometimes, observation will be the only method available to the researcher.

Types of Observation Methods

There are several methods of observation of which any one or a combination of some of them could be used by the observer. Some of these are:

- Structured or unstructured method
- Disguised or undisguised method
- Direct-indirect observation
- Human-mechanical observation

Structured-Unstructured Observation

Whether the observation should be structured or unstructured depends on the data needed. *Example:* A manager of a hotel wants to know "how many of his customers visit the hotel with their families and how many come as single customers. Here, the observation is structured, since it is clear "what is to be observed". He may instruct his waiter store cord this. This information is required to decide

requirements of the chairs and tables and also the ambience.

Disguised-Undisguised Observation

In disguised observation, the respondents do not know that they are being observed. In non- disguised observation, the respondents are well aware that they are being observed. In disguised observation, observers often pose as shoppers. They are known as "mystery shoppers". They are paid by research organizations. The main strength of disguised observation is that it allows for registering the true of the individuals.

Direct-Indirect Observation

In direct observation, the actual behavior or phenomenon of interest is observed. In indirect observation, the results of the consequences of the phenomenon are observed. Suppose, a researcher is interested in knowing about the soft drinks consumption of a student in a hostel room. He may like to observe empty soft drink bottles dropped into the bin. Similarly, the observer may seek the permission of the hotel owner to visit the kitchen or stores. He may carry out a kitchen/stores audit, to find out the consumption of various brands of spice items being used by the hotel. It may be noted that the success of an indirect observation largely depends on "how best the observer is able to identify physical evidence of the problem under study".

Human-Mechanical Observation

Most of the studies in marketing research are based on human observation, wherein trained observers are required to observe and record their observation. In some cases, mechanical devices such as eye cameras are used for observation. One of the major advantages of electrical/ mechanical devices is that their recordings are free from any subjective bias.

Advantages of Observation Method

1. Original data can be collected at the time of occurrence of the event.

2. Observation is done in natural surroundings. Therefore, the facts emerge more clearly, whereas in a questionnaire, experiments have environmental as well as time constraints.

3. Sometimes, the respondents may not like to part with some of the information. Such information can be obtained by the researcher through observation. Observation can also be done on those who cannot articulate.

4. Any bias on the part of the researcher is greatly reduced in the observation method.

Limitations of Observation Method

1. The observer might wait for longer period at the point of observation. And yet the desired event may not take place. Observation is required over a long period of time and hence may not occur.

2. For observation, an extensive training of observers is required.

3. This is an expensive method.

4. External observation provides only superficial indications. To delve beneath the surface is very difficult. Only over behavior can be observed.

5. Two observers may observe the same event, but may draw different inferences.

6. It is very difficult together information on (1) Opinions (2) Intentions.

2.2.2 Interview method

There are different methods of it and which are following:-

Personal Interviews

An interview is called personal when the Interviewer asks the questions face-to-face with the Interviewee. Personal interviews can take place at home, at a shopping mall, on the street, and so on.

Advantages

- The ability to let the Interviewee see, feel and/or taste a product.
- The ability to find the target population. For example, you can find people who have seen a film much more easily outside a theater in which it is playing than by calling phone numbers at random.
- Longer interviews are sometimes tolerated. Particularly with in-home interviews that have been arranged in advance. People may be willing to talk longer face-to-face than to someone on the phone.

Disadvantages

- Personal interviews usually cost more per interview than other methods.
- Change in the characteristics of the population might make sample non-representative.

Telephone Surveys

It is a process of collecting information from sample respondents by calling them over telephone. Surveying by telephone is the most popular interviewing method.

Advantages

- People can usually be contacted faster over the telephone than with other methods.
- You can dial random telephone numbers when you do not have the actual telephone numbers of potential respondents.
- Skilled interviewers can often invite longer or more complete answers than people will give on their own to mail, e-mail surveys.

Disadvantages

- Many telemarketers have given legitimate research a bad name by claiming to be doing research when they start a sales call.
- The growing number of working women often means that no one is at home during the day. This limits calling time to a "window" of about 6-9 p.m. (when you can be sure to interrupt dinner or a favorite TV program).
- You cannot show sample products by phone.

Computer Direct Interviews

These are methods in which the respondents key in(enter)their answers directly in to a computer.

Advantages

- It eliminates data entry and editing costs.
- Answers are more accurate to sensitive questions through a computer than to a person or paper questionnaire.
- Interviewer bias is eliminated. Different interviewers can ask questions in different ways, leading to different results. The computer asks the questions the same way every time.

Disadvantages

- The interviewees must have access to a computer or it must be provided for them.

- As with mail surveys, computer direct interviews may have serious response rate problems in populations due to literacy levels being low.

E-mail Surveys

Email Questionnaire is a new type of questionnaire system that revolutionizes the way on-line questionnaires are conducted. Unlike other on-line questionnaire systems that need a web server to construct, distribute and manage results, Email Questionnaire is totally email based. It works with the existing email system making on-line questionnaire surveys available to anyone with an Internet connection.

Advantages

- Speed: An email questionnaire can gather several thousand responses within a day or two.
- There are practically no costs involved once the setup has been completed.
- Pictures and sound files can be attached.
- The novelty element of an email survey often stimulates higher response levels than ordinary mail surveys.

Disadvantages

- Researcher must possess or purchase a list of email addresses.
- Some people will respond several times or pass questionnaires along to friends to answer.
- Many people dislike unsolicited email even more than unsolicited regular mail.
- Findings cannot be generalized with email surveys. People who have email are different from those who do not, even when matched on demographic characteristics, such as age and gender.
- Email surveys cannot automatically skip questions or randomize question.

Internet/Intranet (Web Page) Survey

Web surveys are rapidly gaining popularity. They have major speed, cost, and flexibility advantages, but also significant sampling limitations. These limitations restrict the groups that can be studied using this technique.

Advantages

- Web page surveys are extremely fast. A questionnaire posted on a popular Web site can gather several thousand responses within a few hours. Many people who will respond to an email invitation to take a Web survey will do so the first day, and most will do so within a few days.
- There is practically no cost involved once the set up has been completed.
- Pictures can be shown. Some Web survey software can also show video and play sound.
- Web page questionnaires can use complex question skipping logic, randomizations and other features which is not possible with paper questionnaires. These features can assure better data.
- Web page questionnaires can use colors, fonts and other formatting options not possible in most email surveys.
- A significant number of people will give more honest answers to questions about sensitive topics, such as drug use or sex, when giving their answers to a computer, instead of to a person or on paper.
- On an average, people give longer answers to open-ended questions on Web page questionnaires than they do on other kinds of self-administered surveys.

Disadvantages

- Current use of the Internet is far from universal. Internet surveys do not reflect the population as a whole. This is true even if a sample of Internet users is selected to match the general population in terms of age, gender and other demographics.
- People can easily quit in the middle of a questionnaire. They are not as likely to complete along questionnaire on the Web as they would be if talking with a good interviewer.
- Depending on your software, there is often no control over people responding multiple times to bias the results.

Mail Questionnaire

Mail questionnaire is a paper questionnaire, which is sent to selected respondents to fill and post filled questionnaire back to the researcher.

Advantages

1. Easier to reach a larger number of respondents throughout the country.

2. Since the interviewer is not present face to face, the influence of interviewer on the respondent is eliminated.
3. This is the only kind of survey you can do if you have the names and addresses of the target population, but not their telephone numbers.
4. Mail surveys allow the respondent to answer at their leisure, rather than at the often inconvenient moment they are contacted for a phone or personal interview. For this reason, they are not considered as intrusive as other kinds of interviews.
5. Where the questions asked are such that they cannot be answered immediately, and needs some thinking on the part of the respondent, the respondent can think over leisurely and give the answer.
6. Saves cost (cheaper than interview).
7. No need to train interviewers.
8. Personal and sensitive questions are well answered in this method.
9. The questionnaire can include pictures - something that is not possible over the phone.

Limitations

1. It is not suitable when questions are difficult and complicated. Example, Do you believe in value price relationship?
2. When the researcher is interested in a spontaneous response, this method is unsuitable. Because thinking time allowed to the respondent will influence the answer.
3. In case of a mail questionnaire, it is not possible to verify whether the respondent himself/ herself has filled the questionnaire. If the questionnaire is directed towards the housewife, say, to know her expenditure on kitchen items, she alone is supposed to answer it. Instead, if her husband answers the questionnaire, the answer may not be correct.
4. Any clarification required by the respondent regarding questions is not possible.
5. If the answers are not correct, the researcher cannot probe further.
6. Poor response (30%) - Not all will reply.
7. In populations of lower educational and literacy levels, response rates to mail surveys are often too small to be useful.

2.2.3 Questionnaire

A questionnaire is a research instrument consisting of a series of questions and other prompts for the purpose of gathering information from respondents. The questionnaire was invented by Sir Francis Galton.

Characteristics of Questionnaire

1. It must be simple. The respondents should be able to understand the questions.
2. It must generate replies that can be easily be recorded by the interviewer.
3. It should be specific, so as to allow the interviewer to keep the interview to the point.
4. It should be well arranged, to facilitate analysis and interpretation.
5. It must keep the respondent interested throughout.

Process of Questionnaire Designing

The following are the seven steps involved in designing a questionnaire:

Step 1: Determine What Information is required

The first question to be asked by the market researcher is "what type of information does he need from the survey?" This is valid because if he omits some information on relevant and vital aspects, his research is not likely to be successful. On the other hand, if he collects information which is not relevant, he is wasting his time and money.

At this stage, information required, and the scope of research should be clear. Therefore, the steps to be followed at the planning stage are:

1. Decide on the topic for research.
2. Get additional information on the research issue, from secondary data and exploratory research. The exploratory research will suggest "what are the relevant variables?"
3. Gather what has been the experience with similar study.

Step 2: Different Types of Questionnaire

1. Structured and Non-disguised
2. Structured and Disguised

3. Non-structured and Disguised
4. Non-structured and Non-disguised

Structured and Non-disguised Questionnaire: Here, questions are structured so as to obtain the facts. The interviewer will ask the questions strictly in accordance with the prearranged order. For example, what are the strengths of soap A in comparison with soap B?

(a) Cost is less
(b) Lasts longer
(c) Better fragrance
(d) Produces more lather

1. Structured and non-disguised questionnaire is widely used in market research. Questions are presented with exactly the same wording and same order to all respondents. The reason for standardizing the question is to ensure that all respondents reply the same question. The purpose of the question is clear. The researcher wants the respondent to choose one of the five options given above.

Example:"Subjects attitude towards Cyber laws and the need for government legislation to regulate it".

Certainly, not needed at present Certainly not needed

I can't say

Very urgently needed

Not urgently needed

2. **Structured and disguised Questionnaire:** This type of questionnaire is least used in marketing research. This type of questionnaire is used to know the peoples' attitude, when a direct undisguised question produces a bias. In this type of questionnaire, what comes out is "what does the respondent know" rather than what he feels. Therefore, the endeavor in this method is to know the respondent's attitude.

Currently, the "Office of Profit" Bill is:

(a) In the Lok Sabha for approval.
(b) Approved by the Lok Sabha and pending in the Rajya Sabha.
(c) Passed by both the Houses, pending the presidential approval.

(d) The bill is being passed by the President.

Depending on which answer the respondent chooses, his knowledge on the subject is classified.

In a disguised type, the respondent is not informed of the purpose of the questionnaire. Here the purpose is to hide "what is expected from the respondent?"

Example: "Tell me your opinion about Mr. Ben's healing effect show conducted at Bangalore?"

"What do you think about the Babri Masjid demolition?"

3. **Non-Structured and Disguised Questionnaire:** The main objective is to conceal the topic of enquiry by using a disguised stimulus. Though the stimulus is standardized by the researcher, the respondent is allowed to answer in an unstructured manner. The assumption made here is that individual's reaction is an indication of respondent's basic perception. Projective techniques are examples of non-structured disguised technique. The techniques involve the use of a vague stimulus, which an individual is asked to expand or describe or build a story, three common types under this category are(a) Word association (b)Sentence completion (c) Storytelling.

4. **Non-structured and Non disguised Questionnaire*:*** Here the purpose of the study is clear, but the responses to the question are open-ended. Example: "How do you feel about the Cyber law currently in practice and its need for further modification"? The initial part of the question is consistent. After presenting the initial question, the interview becomes much unstructured as the interviewer probes more deeply. Subsequent answers by the respondents determine the direction the interviewer takes next. The question asked by the interviewer varies from person to person. This method is called "the depth interview". The major advantage of this method is the freedom permitted to the interviewer. By not restricting the respondents to a set of replies, the experienced interviewers will be above to get the information from the respondents fairly and accurately.

Step 3: Type of Questions

Open-ended Questions

These are questions where respondents are free to answer in their own words. Example: "What factor do you consider while buying a suit"? If multiple choices are given, it could be color, price, style, brand, etc., but some respondents may mention attributes which may not occur to the researcher. Therefore, open-ended questions are useful in exploratory research, where all possible alternatives are explored.

The greatest disadvantage of open-ended questions is that the researcher has to note down the answer of the respondents verbatim. Therefore, there is a likelihood of the researcher failing to record some information. Another problem with open-ended question is that the respondents may not use the same frame of reference.

Example: "What is the most important attribute in a job?"

Ans: Pay

The respondent may have meant "basic pay" but interviewer may think that the respondent is talking about "total pay including dearness allowance and incentive". Since both of them refer to pay, it is impossible to separate two different frames.

Dichotomous Question

These questions have only two answers, 'Yes' or 'no' , 'true' or 'false' 'use' or 'don't use'. Do you use toothpaste? Yes …….. No……

There is no third answer. However sometimes, there can be a third answer:

Example: "Do you like to watch movies?"

Ans: Neither like nor dislike.

Dichotomous questions are most convenient and easy to answer. A major disadvantage of dichotomous question is that it limits the respondent's response. This may lead to measurement error.

Close-Ended Questions

There are two basic formats in this type:

- Make one or more choices among the alternatives.
- Rate the alternatives.

Closed-ended questionnaires are easy to answer. It requires less effort on the part of the interviewer. Tabulation and analysis is easier. There are lesser errors, since the same questions are asked to everyone. The time taken to respond is lesser. We can compare the answer of one respondent to another respondent.

Step 4: Wordings of Questions

Wordings of particular questions could have a large impact on how the respondent interprets them. Even a small shift in the wording could alter the respondent's answer.

Example: "Don't you think that Brazil played poorly in the FIFA cup?" The answer will be 'yes'. Many of them, who do not have any idea about the game, will also most likely say 'yes'. If the question is worded in a slightly different manner, the response will be different.

Example:"Do you think that, Brazil played poorly in the FIFA cup?"This is a straight forward question. The answer could be 'yes', 'no' or 'don't know' depending on the knowledge the respondents have about the game.

Step 5: Sequence and Layout

Some guidelines for sequencing the questionnaire are as follows:

Divide the questionnaire into three parts:

1. Basic information
2. Classification
3. Identification information.

Items such as age, sex, income, education, etc., are questioned in the classification section. The identification part involves body of the questionnaire. Always move from general to specific questions on the topic. This is known as funnel sequence.

Layout: How the questionnaire looks or appears.

Example: Clear instructions, gaps between questions, answers and spaces are part of layout. Two different layouts are shown below:

Layout - 1 How old is your bike?

.....Less than1year.....1to2years.....2to4years more than 4years.

Layout - 2 how old is your bike?

..... Less than 1 year

..... 1 to 2 years.

.....2 to 4 years.

..... More than 4 years.

From the above example, it is clear that layout - 2 is better. This is because likely respondent error due to confusion is minimized.

Therefore, while preparing a questionnaire start with a general question. This is followed by a direct and simple question. This is followed by more focused questions .This will elicit maximum information.

Step 6: Pretesting of Questionnaire

Pretesting of a questionnaire is done to detect any flaws that might be present. For example, the word used by researcher must convey the same meaning to the respondents. Are instructions clear skip questions clear? One of the prime conditions for pretesting is that the sample chosen for pretesting should be similar to the respondents who are ultimately going to participate. Just because a few chosen respondents fill in all the questions going does not mean that the questionnaire is sound.

How Many Questions to be asked? The questionnaire should not be too long as the response will be poor. There is no rule to decide this. However, the researcher should consider that if he were the respondent, how he would react to a lengthy questionnaire. One way of deciding the length of the questionnaire is to calculate the time taken to complete the questionnaire. He can give the questionnaire to a few known people to seek their opinion.

Step 7: Revise and Preparation of Final Questionnaire

Final questionnaire may be prepared after pre testing the questionnaire with the small group of respondents. Questionnaire should be revised for the following:

i. To correct the spellings.

ii. To place the questions in proper order to avoid the contextual bias.

iii. To remove the words which are not familiar to respondents?

iv. To add or remove questions arise in the process of pretest, If any.

v. Top urge the words with double meaning, etc.

2.4 Secondary Data Collection Method

In research, secondary data is collecting and possibly processed by people other than the researcher in question. Common sources of secondary data for social science include censuses, large surveys, and organizational records. In sociology primary data is data you have collected yourself and secondary data is data you have gathered from primary sources to create new research. In terms of historical research, these two terms have different meanings. A primary source is a book or set of archival records .A secondary source is a summary of a book or set of records. Secondary data are statistics that already exist. They have been gathered not for immediate use. This may be described as "those data that have been compiled by some agency other than the user". Secondary data can be classified as:

Internal Secondary Data:-

Internal secondary data is a part of the company's record, for which research is already conducted. Internal data are those that are found within the organization. *Example:* Sales in units, credit outstanding, call reports of sales persons, daily production report, monthly collection report, etc.

External Secondary Data:-

The data collected by the researcher from outside the company. This can be divided into four parts:

1. Census data
2. Individual project report being published
3. Data collected for sale on a commercial basis called syndicated data
4. Miscellaneous data

The following are some of the data that can obtain by census records:

- Census of the wholesale trade
- Census of the retail trade
- Population Census
- Census of manufacturing industries
- Individual project report being published
- Encyclopedia of business information sources
- Product finder

- Thomas registers etc.

Benefits and Limitations of Secondary Data

Benefits

It is far cheaper to collect secondary data than to obtain primary data. For the same level of research budget a thorough examination of secondary sources can yield a great deal of information than can be had through a primary data collection exercise. The time involved in searching secondary sources is much less than that needed to complete primary data collection.

Secondary sources of information can yield more accurate data than that obtained through primary research. This is not always true but where a government or international agency has undertaken a large scale survey, or even a census, this is likely to yield far more accurate results than custom designed and executed surveys when these are based on relatively small sample sizes.

It should not be forgotten that secondary data can play a substantial role in the exploratory phase of the research when the task at hand is to define the research problem and to generate hypotheses. The assembly and analysis of secondary data almost invariably improve the researcher's understanding of the marketing problem, the various lines of inquiry that could or should be followed and the alternative courses of actions which might be pursued.

Secondary sources help define the population. Secondary data can be extremely useful both in defining the population and in structuring the sample to be taken. For instance, government statistics on a country's agriculture will help decide how to stratify a sample and, once sample estimates have been calculated, these can be used to project those estimates to the population.

Limitations

1. *Definition:* The researcher, when making use of secondary data, may misinterpret the definitions used by those responsible for its preparation and drawer roneous conclusions
2. *Measurement error:* When a researcher conducts fieldwork she/he is possibly able to estimate inaccuracies in measurement through the standard deviation and standard error, but these are sometimes not published in secondary sources. The problem is sometimes not so much 'error' but differences in the levels of accuracy required by decision makers.
3. *Sourcebias:* Researchers face the problem of vested interests when they consult secondary

sources. Those responsible for their compilation may have reasons for wishing to present a more optimistic or pessimistic set of results for their organization i.e., exaggerated figures or inflated estimates may be stated.

4. *Reliability:* There liability of published statistics may vary over time. Because the systems of collecting data or geographical or administrative boundaries may be changed, or the basis for stratifying a sample may have altered. Other aspects of research methodology that affect the reliability of secondary data is the sample size, response rate, questionnaire design and modes of analysis without any indication of this to the reader of published statistics.
5. *Time scale:* The time period during which secondary data was first compiled may have a substantial effect upon the nature of the data for example: Most censuses take place atten- year intervals, so data from this and other published sources may be out-of-date at the time there searcher wants to make use of the statistics.

2.5 Check Your Progress

1. A major disadvantage of dichotomous question is that it----------the respondent's response.
2. Open-ended questions are useful in ..... research, where all possible alternatives are explored.
3. Internal secondary data is a part of the..... ……record.
4. External Secondary Data can be divided into ……..parts.
5. Internal data are those that are found…….the organization.

2.6 Summary

Primary data may pertain to life style, income, awareness or any other attribute of individuals or groups. There are mainly two ways of collecting primary data namely: (a) Observation (b) By questioning the appropriate sample. Observation method has a limitation i.e., certain attitudes, knowledge, motivation, etc. cannot be measured by this method. For this reason , researcher needs to communicate. Communication method is classified based on whether it is structured or disguised. Questionnaire is easy to administer. This type is most suited for descriptive research. If the researcher wants to do exploratory sturdy, unstructured method is better. In unstructured method questions will

have to be framed based on the answer by the respondent. Questionnaire can be administered either in person or online or Mail questionnaire. Each of these methods has advantages and disadvantages. Questions in a questionnaire may be classified into (a) Open question (b) Close ended questions (c) Dichotomous questions, etc. While formulating questions, care has to be taken with respect to question wording, vocabulary; leading, loading and confusing questions should be avoided. Further it is desirable that questions should not be complex, or too long. It is also implied that proper sequencing will enable the respondent to answer the question easily. The researcher must maintain a balanced scale and must use a funnel approach. Pretesting of the questionnaire is preferred before introducing to a large population. Secondary data are statistics that already exists. Secondary data may not be readily used because these data are collected for some other purpose. Secondary data has its own advantages and disadvantages. There are two types of secondary data (1) Internal and (2) External secondary data. Census is the most important among secondary data. Syndicated data is an important form of secondary data. Syndicated data may be classified into (a) Consumer purchase data (b) Retailer and wholesaler data (c) Advertising data. Each has advantages and disadvantages.

2.7 Keywords

1. **Dichotomous question:** These questions have only two answers, like 'yes or no'.

2. **Disguised observation:** The observation under which the respondents do not know that they are being observed.

3. **Non-Disguised observation:** The observation in which the respondents are well aware that they are being observed.

4. **External Data:** The data collected by the researcher from outside the company.

5. **Internal Data:** Internal data are those that are found within the organization.

6. **Panel Type Data:** This is one type of syndicated data in which there are consumer panels.

7. **Secondary Data:** Secondary data is collecting and possibly processed by people other than the researcher in question.

8. **Syndicated Data:** Data collected by this method is sold to interested clients on payment.

2.8 Self- Assessment Test

Q1. What is primary data?

Q2. What are the various methods of collecting primary data?

Q3. What is questionnaire? What are its importance and characteristics?

Q4. Explain open ended and close ended questions in a questionnaire.

Q5. What are the advantages and disadvantages of primary data?

Q6. What are the sources of secondary data?

Q7. What are the types of secondary data?

Q8. What are the advantages and disadvantages of secondary data?

2.9 Answers to check Your Progress

1. Limits

2. Exploratory

3. Company's

4. Four

5. Within

2.10 References/ Suggested Readings

Gupta, S. P.: Statistical Methods, Sultan chand and sons, New Delhi.

Kumar, S.: Practical Statistics, Sultan chand and sons, New Delhi

Levin, R. and David, S. R.: Statistics for Management, Prentice Hall, New Delhi.

Gupta, C. B.: Introduction to Statistical Methods, Ram Prashad, New Delhi.

Sancheti, D. C. and Kapoor, V. K.: Business Statistics.

Agarwal, B. L.: Basic Statistics, New age International.

Kapur, S. K.: Elements of Practical Statistic, Oxford & IBH Publishers.

Subject : Business Statistics-1	
Course Code : BCOM 302	Author : Mr Ankit
Lesson No. : 3	Vetter: Prof. Suresh Kumar Mittal

CLASSIFICATION AND TABULATION OF DATA

STRUCTURE

3.10 References/ Suggested Readings

3.0 LEARNING OBJECTIVES

After going through this lesson, the learner should be able to know:

- Meaning, Functions, Rules and Basis of Classification of Data
- Concept and Types of Frequency Distribution
- Concept, Types, Components and Requisites of Tabulation
- Types and General rules of Presentation of Data
- Graphical Representation of Frequency Distribution

3.1 INTRODUCTION

In the previous lesson, the author discussed about the concept of data and various techniques of gathering data for statistical investigation. Collection of data without any modification is called raw data. The raw data is highly disorganized data. It is very large in amount and difficult to handle. It is very difficult and tedious task to make a meaningful conclusion from raw data for any statistical enquiry. Hence, for any statistical analysis, proper arrangement of raw data is required. For example, find out the overall result of 90 undergraduates in Statistics, we collect their marks and present them through Table 3.0 as shown below.

TABLE 3.0 OVERALL RESULTS OF 90 UNDERGRADUATES IN STATISTICS

47	46	56	60	61	62	58	39	28
60	44	25	35	65	54	71	73	49
55	49	29	40	72	43	48	47	53
35	35	31	33	38	82	83	57	90
43	38	36	29	48	39	57	37	67
44	20	70	65	38	44	50	41	66
89	90	68	29	21	62	46	48	64
59	70	56	54	52	84	66	61	61

70	27	15	26	64	48	78	62	63
53	19	29	30	78	88	82	65	40

As you see in Table 3.0, the numbers in Table 3.0 are not arranged in any order. To make any statistical inferences from it, firstly arrange the marks either in ascending or descending order. It is already a very difficult job and if the strength of students is increased from ninety to one thousand, it will become even more difficult. Therefore, the next important step is to organize and edit the raw data to show it in an easily understandable summarized form. It highlights all key features, facilitates comparisons and makes it appropriate for analysis and interpretations. Following are the three ways to organize the data:

- Classification of data
- Tabulation of data
- Presentation of data

3.2 CLASSIFICATION OF DATA

This method of classification means ordering raw data into groups or classes as per their similarities. Some important definitions of classification are:

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

According to '**Tuttle A.M.**' – "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 the category".

Thus, classification is systematic organization of the raw data into various classes depending upon the nature, scope and objectives of the enquiry. When facts of related features are placed in one class, this allows tracing them easily, making comparison and drawing conclusions without any trouble.

3.2.0 FUNCTIONS OF CLASSIFICATION

1. **Condensation of Data:** Classification changes the large amount of unmanageable raw data in a summarized or condensed form. It makes raw data freely understandable and also highlights the major features of the data.

2. **Facilitate Comparison:** Classification is helpful in making a meaningful comparison between different fields of a given data. For example, classification of population of a country as per their income level which helps us to study the standard of living of people of that country.
3. **Study the Relationship:** The classification of the given data on the basis of two or more criteria enables us to study the relationship between them.
4. **Facilitate Statistical Treatment:** The organization of huge incomparable data into relatively comparable groups or classes on the basis of their similarities presents homogeneity among diversity. This makes data more logical, useful and responsive for further processing.

3.2.1 RULES OF CLASSIFICATION

It is important method of organization of data before tabulation and presentation. Although, a good classification depends on the nature of data and objectives of statistical enquiry, no hard and fast rule is laid down for it. However, following are the general rules for a good classification of data:

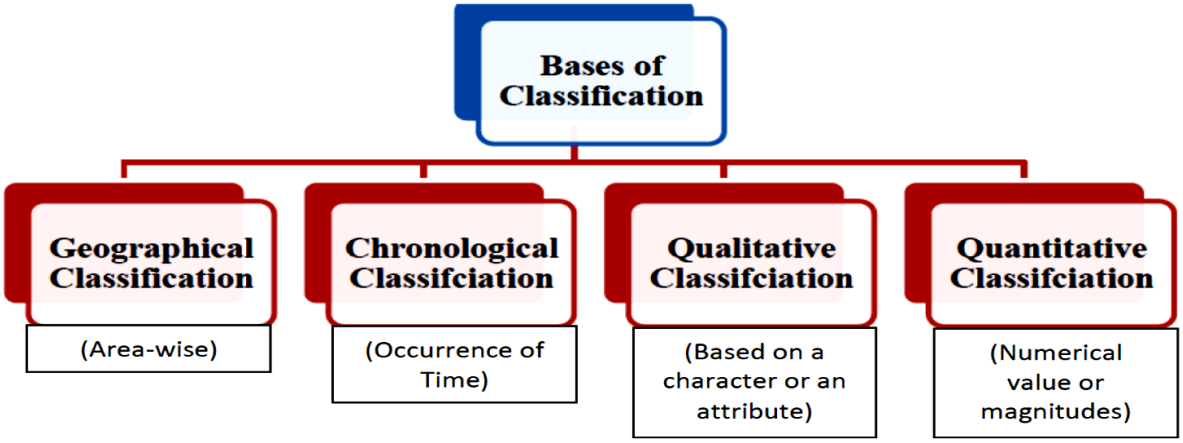
1. **It must be clearly defined:** The classes must be strictly definite and should not show any ambiguity or uncertainty. In other words, there should not be any doubt or confusion regarding the location of the observations in the given classes. For example, if we have to classify a group of individuals as 'employed' and 'un-employed' or 'literate' and 'illiterate', it is imperative to define it in clear cut terms as to what we mean by an employed and unemployed person; a literate and illiterate person.
2. **It must be comprehensive and mutually exclusive:** According to this**,** the classification should be exhaustive in the sense that each and every item in the data must belong to one of the classes. A good classification should be free from the residual class like 'others' or 'miscellaneous' because such classes do not reveal the characteristics of the data completely. However, if the classes are very large in number, it becomes necessary to introduce this 'residual class', otherwise the purpose of classification viz., condensation of the data will be defeated. Further, the various classes should be mutually disjoint or non-overlapping so that an observed value belongs to one and only one of the classes. For instance, if we classify the students in a college by sex *i.e.*, as males and females, the two classes are mutually exclusive. But if the same group is classified as males, females and addicts to a particular drug then the classification is faulty because the group "addicts to a particular drug" includes both males and females.

3. **It must be steady:** According to this, a classification is best when it is well established and constant. It means, the outline of classification must remain same throughout the analysis to draw a meaningful comparison of results. Adopting different classification techniques in the same analysis may lead to misleading interpretation.
4. **It must be suitable for the purpose:** According to this, knowing the analysis objective while classifying the data is important. We must avoid any such classification of data which does not suit the enquiry purpose. For example, to know the nexus between the university education and sex, classification of students based on their age and religion is irrelevant.
5. **It must be flexible:** According to this rule, classification of data should be done in such a way which enables us to do any modification in future. Due to constant updates, a change in the statistical methods of classifications may be required. Therefore, it is necessary for the classification to be flexible and adjustable to the changed situations.

3.2.2 BASIS OF CLASSIFICATION

Classification can be done in many ways. For example, books may be classified according to subjects like "History", "Geography", "Mathematics", "Science", and so on. They may also be classified according to author name in an alphabetical order or as per their publication year. Similarly, the raw data should also be classified based on the purpose and nature of statistical enquiry. The figure 3.0 portrays the various bases of classification of raw data:

FIGURE 3.0 BASIS OF CLASSIFICATION OF DATA



Geographical Classification

Classification based on geographical site such as countries, states, districts, regions, zones, areas etc., is called geographical or spatial classification. It is usually presented either in an alphabetical order or conferring to a size or value which highlights an important area or region. Table 3.1 portraying yield of rice per hectare of different nations in descending order is a very good example of geographical classification.

TABLE 3.1 YIELD OF WHEAT PER HECTARE OF DIFFERENT NATIONS

Nation	Yield of rice (kg per Hectare)
UAR	750·5
USSR	735·5
Syria	620·2
USA	580
Sudan	349·8
China	270·2
Pakistan	260·5
India	125

Chronological Classification:

The classification in which the data is arranged either in ascending or descending order based on the time gap between several years, quarters, months, weeks, days etc. is defined as chronological classification. Many examples of chronological classification are seen in real life situations like the yearly production of a business concern; the sales of a company over several years; the population of Asia for years and so on. Generally, in economics and business statistics, chronological classification is used in the form of time series data. Table 3.2 shows the population of China for several decades in descending order.

TABLE 3.2 POPULATION OF CHINA FOR DIFFERENT DECADES

Year	Population (In crores)
2011	121.0
2001	102.7
1991	81.8
1981	68.4
1971	54.6
1961	43.8
1951	35.7

Qualitative Classification:

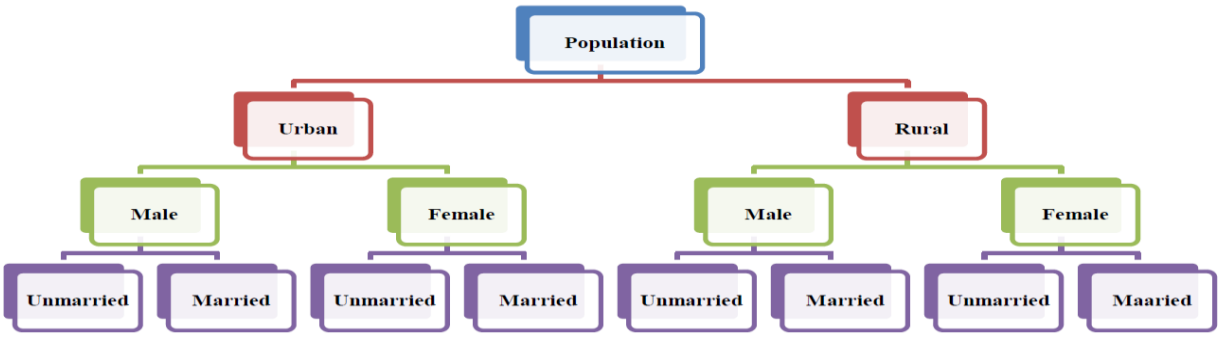
When the data is classified on the basis of presence or absence of some qualities or attributes which are not measurable quantitatively, it is termed as qualitative classification. These qualities are employment, occupation, IQ level, sex, nationality, gender, religion, marital status and so on. Types of qualitative classification are as follows:

- **Simple or Dichotomous classification:** When the data is classified on the basis of only presence or absence of an attribute, it is termed as simple or dichotomous classification. For instance, classifying the employees of a company as truthful or untruthful; man or woman; working or non-working; attractive or not attractive and so on. Figure 3.1 shows the simple or dichotomous classification of employees of a company on gender basis.

FIGURE 3.1 SIMPLE OR DICHOTOMOUS CLASSIFICATIONS OF THE EMPLOYEES OF A COMPANY ON GENDER BASIS

- **Manifold Classification:** If the data is classified into different classes for one or more attributes and then subdivided into sub classes also, it is termed as manifold classification. Figure 3.2 shows the population is firstly divided on the basis of area (Urban and Rural), then subdivided on the basis of gender (Male and Female) and again as per marital status (Married and Unmarried).

FIGURE 3.2 MANIFOLD CLASSIFICATION OF POPULATION ON GENDER AND MARITAL STATUS BASIS



Quantitative Classification:

When the classification is done based on those attributes or qualities of a phenomenon which is quantitatively measurable, it is defined as quantitative classification. Some of the quantitative attributes are price, production, income, expenditure, sales, profit, age, height, weight etc. It is also called classification by variables because the quantitative attribute under study is also known as variable. For example, Table 3.3 portrays the daily earning of 60 departmental stores.

TABLE 3.3 DAILY EARNING OF 60 DIFFERENT STORE

Daily Earning (In Thousand rupee)	Number of Stores
0 – 100	6
101 – 200	14
201 – 300	8
301 – 400	10
401 – 500	8
501 – 600	6

601 – 700	4
701 – 800	4

Variables are of two types:-

- Continuous Variable
- Discrete Variable

Continuous Variable: It takes all possible values in whole and also in fraction in a given range. For example, the body temperature of a person can be depicted in all possible values to the nearest fraction of digit on different scales like degree of Celsius, degree of Fahrenheit and so on. The attributes like height, weight and distance are also some of the examples of continuous variables. Therefore, if a variable has the ability of jumping from one value to another by infinitely small degrees, it becomes continuous variable.

Discrete variable: It is a variable which is not capable of taking all the possible values within a given defined range. For instance, the number of males and females out of 100 selected candidates in an interview is an example of discrete variable, since in this case number of males and females can vary only from 0 to 100. The attributes like family size, residents of a country, number of mishaps on the highway etc., are some other examples of discrete variables. In simple words, values of a discrete variable can take only finite "jumps" like 1, 2, 3, etc.

3.3 CONCEPT OF FREQUENCY DISTRIBUTION

A frequency distribution is an extensive way to represent the data of a quantitative variable. The frequency distribution table of a variable shows its different values categorized in various classes with their matching class frequencies.

3.3.0 IMPORTANT TERMS OF FREQUENCY DISTRIBUTION

Table 3.4 shows the important terms used in a frequency distribution:

TABLE 3.4: IMPORTANT TERMS OF FREQUENCY DISTRIBUTION

Class	Frequency	Lower Class Limit	Upper Class Limit	Class Mid-point

0 – 10	1	0	10	5
10 – 20	8	10	20	15
20 – 30	6	20	30	25
30 – 40	7	30	40	35
40 – 50	21	40	50	45
50 – 60	23	50	60	55
60 – 70	19	60	70	65
70 – 80	6	70	80	75
80 – 90	5	80	90	85
90 – 100	4	90	100	95

- **Class: Grouping of values of raw data to facilitate formation of its frequency distribution is called a class.** Usually, the number of classes varies from six to fifteen. When equal sized class intervals are used, number of classes can be the calculated by dividing the difference between the largest and the smallest values of variable i.e. range by the size of the class intervals.

$$\textbf{Number of Class} = \frac{\textbf{Largest value of variable} - \textbf{smallest value of variable}}{\textbf{Size of class interval}}$$

- **Class Frequency:** Frequency is a measure of how often something has happened. The frequency of any observation tells you that how many times a specific observation has occurred in an observed data. In other words, class frequency refers to the number of values in a particular class. It can be counted through tally marks against every class. Table 3.5 shows counting of the class frequency by tally method.

TABLE 3.5 COUNTING OF CLASS FREQUENCY BY TALLY METHOD

Class Interval	Frequency	Tally Bar
10 – 20	4	*IIII*

20 – 30	3	*III*
30 – 40	2	*II*
40 – 50	1	*I*
50 – 60	5	*IIII*
60 – 70	6	*IIII I*

- **Class Limit:** The two ends of a class are class limits. The lowest value of the class is called lower class limit and the highest one is upper class limit. For example, the class limits for the class: 40–50 are 40 and 50. 40 is the lower class limit while 50 is the upper class limit.
- **Class Interval or Width:** The difference between the upper and the lower class limit is called class interval or class width. For class 40–50, the class interval is 10. Open-ended classes like "70 and over" or "less than 10" are generally not recommended. The upper and lower class limits should be fixed in such a way that frequency of each class concentrates towards mid value of the each class interval. Class interval is of three types:-

  Inclusive Class Interval: In case of inclusive class interval, values of both lower and upper limits of a class are included in the frequency of that class. In this method, the lower limit of a class is not included in the upper limit of the preceding class. The inclusive class interval has to be changed to an exclusive one for the analysis of frequency distribution. For this conversion, the value of 0.5 should be subtracted from the lower class limit and added to the upper class limit. The following formula is used in the conversion of an inclusive class interval into an exclusive one.

$$\textbf{Adjusted Class Interval} = \frac{\textbf{Upper limit of first class} - \textbf{Lower limit of second class}}{\mathbf{2}}$$

Table 3.6 shows an example of conversion of an inclusive interval into exclusive interval.

TABLE 3.6 ADJUSTED CLASS INTERVALS

Inclusive Class		Adjusted Class	Frequency
10 – 19	19–20/2 = -0.5	9.5 – 19.5	4
20 – 29	29-30/2 = -0.5	19.5 – 29.5	3

30 – 39	39-40/2 = -0.5	29.5 – 39.5	2
40 – 49	49-50/2 = -0.5	39.5 – 49.5	1
50 – 59	59-60/2 = -0.5	49.5 – 59.5	5

Exclusive Class Interval: In exclusive class interval, an item of either the upper class limit or the lower one should be excluded from the frequency of that class. Here, the upper limit of one class is equivalent to the lower limit of the following class which ensures continuity between two successive classes. An example of an exclusive class interval is shown in Table 3.7 below.

TABLE 3.7 EXCLUSIVE CLASS INTERVALS

Exclusive Class Intervals	Frequency
100 – 200	4
200 – 300	3
300 – 400	2
400 – 500	1
500 – 600	5
600 – 700	6

Note: Inclusive class intervals are often used for both, continuous as well as discrete variables. While, exclusive class intervals are generally used only for discrete variables.

Open End Class Interval:

When in a class interval series l**ess than or below** is stated in place of the lower limit of first class interval and **more than or above** is mentioned in place of the upper limit of the series, such types of intervals are known as **open ended class intervals.** Table 3.8 is a good example of open end class interval:

TABLE 3.8 OPEN END CLASS INTERVALS

Marks	Frequency

Below 10	4
20 – 30	3
30 – 40	2
40 – 50	1
50 – 60	5
60 and above	6
Total	21

For further statistical inferences, open-ended class intervals should be changed into specified class limits by giving them the same magnitude or class size as that of other class intervals. In the above example, the magnitude of other class intervals is 10. So, the open-end class intervals should be written as 0-10 and 60-70 for first and last class interval of the series.

- **Class Mid-Point:** is the middle value of a class which lies halfway between the lower and the upper class limit of a class. It can be ascertained by the following formula:

$$\textbf{Class Mid point} = \frac{\textbf{Upper class limit} + \textbf{Lower class limit}}{\textbf{2}}$$

Table 3.9 portrays the calculation of mid values of a class intervals series:

TABLE 3.9 CLASS MID-POINTS

Class Interval	Mid Value	Frequency
10 – 20	10 + 20 /2 = 15	4
20 – 30	20 + 30 /2 = 25	3
30 – 40	30 + 40 /2 = 35	2
40 – 50	40 + 50 /2 = 45	1
50 – 60	50 + 60 /2 = 55	5
60 – 70	60 + 70 /2 = 65	6

Sometimes, in a series, mid values are given instead of class intervals. Therefore, you have to change these mid values into class intervals for statistical analysis. You have to calculate the lower and upper limit with mid-point value by applying following formulas:

$$\textbf{Lower Class Limit} = \textbf{Mid Value (M)} - \frac{\textbf{Difference between mid values (i)}}{\textbf{2}}$$

$$\textbf{Upper Class Limit} = \textbf{Mid Value (M)} + \frac{\textbf{Difference between mid values (i)}}{\textbf{2}}$$

Table 3.10 shows the conversion of a mid-value series into a class interval series by applying above formula:

TABLE 3.10 CONVERSION OF MID VALUE SERIES INTO A CLASS INTERVAL SERIES

Mid Value	Frequency	Lower Class Limit	Upper Class Limit	Class Interval
15	4	15 – 10 / 2 = 10	15 + 10 / 2 = 20	10 – 20
25	3	25 – 10 / 2 = 20	25 + 10 / 2 = 30	20 – 30
35	2	35 – 10 / 2 = 30	35 + 10 / 2 = 40	30 – 40
45	1	45 – 10 / 2 = 40	45 + 10 / 2 = 50	40 – 50
55	5	55 – 10 / 2 = 50	55 + 10 / 2 = 60	50 – 60
65	6	65 – 10 / 2 = 60	65 + 10 / 2 = 70	60 – 70

3.3.1 TYPES OF FREQUENCY DISTRIBUTION

Following are the different types of the frequency distribution:

- Ungrouped Frequency Distribution
- Grouped Frequency Distribution
- Cumulative Frequency Distribution
- Relative Frequency Distribution
- Bivariate Frequency Distribution

UNGROUPED FREQUENCY DISTRIBUTION

As the data is expressed in discrete form, this distribution is also called discrete frequency distribution. In this type of distribution, we count the number of times the value of each variable occurring in a dataset. We do not make groups or class intervals in an ungrouped frequency distribution. It shows the frequency of an item as a separate data value instead of group values. Rather, we mention exact frequency of each value. Following examples are helpful in understanding of ungrouped frequency distribution:

Example 3.0: A review of 20 pages of an accounting book shows the following printing mistakes:

0	1	3	1	2	5	3	0	1	2
2	3	3	4	3	0	4	4	1	4

Let's make a frequency distribution of this data.

Solution: Let us assume printing mistake is a discrete variable denoted by X. X can have six values i.e. 0, 1, 2, 3, 4 and 5. So, we have 6 classes, each class having its own frequency value (Number of Pages).

TABLE 3.11 FREQUENCY DISTRIBUTION OF PRINTING MISTAKES (X) OF 30 PAGES OF A BOOK

Printing Mistakes (X)	Tally Bar	Frequency (No. of Pages)
0	III	3
1	IIII	4
2	III	3
3	IIII	5
4	IIII	4
5	I	1
Total		20

Example 3.1: Following data shows the responses of 18 students of a university regarding how many books they read in a year. Make an Ungrouped Frequency Distribution table for this data. The responses are shown below:

7	3	0	7	2	1
4	4	5	6	6	3
3	7	1	1	0	2

Solution: Let us assume the number of books read per year by a student is a discrete variable denoted by Y. It has eight values i.e. 0, 1, 2, 3, 4, 5, 6 and 7. Therefore, we have 8 classes with their respective frequencies shown in Table 3.12.

Table 3.12 UNGROUPED FREQUENCY DISTRIBUTION OF NUMBER OF BOOKS READ PER YEAR BY 18 STUDENTS OF A UNIVERSITY

Books	Tally Bar	Frequency (Number of Students)
0	II	2
1	III	3
2	II	2
3	III	3
4	II	2
5	I	1
6	II	2
7	III	3
Total		18

Example 3.2: Make an ungrouped frequency distribution of marks obtained by 20 students in an exam. The marks are given below:

5	20	20	12	5
5	5	15	12	20
20	15	15	15	15
18	5	18	18	18

Solution: In the above example, marks obtained by students in an exam are the discrete variable (X). Following table shows the results of the above data:

TABLE 3.13 UNGROUPED FREQUENCY DISTRIBUTION OF MARKS OBTAINED BY 20 STUDENTS

Marks (X)	Tally Bar	Number of Students (Frequency)
5	𝍸	5
12	II	2
15	𝍸	5
18	IIII	4
20	IIII	4
Total		20

GROUPED FREQUENCY DISTRIBUTION

It is also known as continuous frequency distribution. It is an apt way of representing the organisation of raw data of a continuous variable. In this type of frequency distribution, the data is first arranged either in ascending or descending order and then divided into groups called class intervals. The frequency of data of each class interval is marked as frequency of that particular class interval. The following examples are helpful in understanding of grouped frequency distribution:

Example 3.3: The data given below is related to marks of 40 candidates in a test for the selection purpose of a job:

41	17	83	63	55	92	60	58	70	06
67	82	33	44	57	49	34	73	54	63
36	52	32	75	60	33	09	79	28	30
42	93	43	80	03	32	57	67	84	64

Take First Class as 0-9. Prepare a frequency distribution table.

Solution:

TABLE 3.14 FREQUENCY DISTRIBUTION OF THE MARKS OF 40 APPLICANTS

Marks	Tally Bar	Frequency
0-9	III	3
10-19	I	1
20-29	I	1
30-39	IIII II	7
40-49	IIII	5
50-59	IIII I	6
60-69	IIII II	7
70-79	IIII	4
80-89	IIII	4
90-99	II	2
Total		40

Example 3.4: Consider the marks obtained by 60 students of 6th standard in an exam. The maximum marks for this exam are 60.

23	18	9	30	22	25	9	41	18	28
8	32	8	10	40	30	15	35	22	30
13	44	25	27	39	24	20	45	20	33
17	27	37	22	17	29	19	31	57	40
11	34	56	46	38	33	55	26	42	32
10	15	22	25	30	35	45	49	39	36

Make a grouped frequency distribution table for the above data.

Solution: We shall make a table with a group of observations say 0 to 15, 15 to 30 and so on.

TABLE 3.15 GROUPED FREQUENCY DISTRIBUTION

Group	Tally Bar	Frequency
0-15	𝍸 III	8
15-30	𝍸 𝍸 𝍸 𝍸 III	23
30-45	𝍸 𝍸 𝍸 𝍸 II	22
45-60	𝍸 II	7
Total		50

CUMULATIVE FREQUENCY DISTRIBUTION

Normally, a frequency distribution provides information regarding the number of times a particular value of a variable (discrete or continuous) occurs. But it does not tell the number of observations having a value "less than" or "more than" a particular value of a variable. This information is provided by 'Cumulative Frequency Distribution'. It is a modified version of simple frequency distribution. It is created by continuously adding the frequencies of the values of a variable as per a certain law. These frequencies are called cumulative frequencies *(c.f.).* It can be of two types:

More than Cumulative Frequency: It can be found by calculating the cumulative total of frequencies starting from the highest value of the variable to the lowest. In 'more than' cumulative frequency distribution, the *c.f.* refers to the lower limit of the corresponding class.

Less than Cumulative Frequency: It can be obtained by adding successively the frequencies of all the previous values including the frequency of variable against which the totals are written. To make a less than cumulative frequency, the values are arranged in ascending order of magnitude. In 'less than' cumulative frequency distribution, the *c.f.* refers to the upper limit of the corresponding class.

Now, look at some examples to understand about cumulative frequency distribution:

Example 3.5: Convert the following table into 'more than cumulative frequency distribution table.

Monthly Salary (Less Than '000)	30	60	90	120	150
Number of Employees	46	93	167	188	200

Solution: Firstly, we have to convert the given data into continuous frequency distribution, as shown in Table 3·16, and then, 'more than' cumulative frequency distribution can be obtained as per Table 3.16 (a):

TABLE 3.16 - CONTINUOUS FREQUENCY DISTRIBUTION

Monthly Salary (in '000 Rs.)	No. of Employees (f)	More than c.f.
0-30	46	46 + (47+74+21+12) = 200
30-60	93 – 46 = 47	47 + (74+21+12) = 154
60-90	167 – 93 = 74	74 + (21+12) = 107
90-120	188 – 167 = 21	21 + 12 = 33
120-150	200 – 188 = 12	12

Table 3.16 (a) - MORE THAN CUMULATIVE FREQUENCY DISTRIBUTION

Employees' salary ('000 Rs.)	No. of Employees
More than 0	200
More than 30	154
More than 60	107
More than 90	33
More than 120	12

Example 3.6: The frequency distribution of scores obtained by 150 candidates in a commerce entrance test is as follows.

Scores	*No. of Candidates*
500 - 550	*35*
550 – 600	*27*
600 – 650	*24*
650 – 700	*17*
700 – 750	*28*
750 – 800	*19*

Make less than and more than cumulative frequency distribution tables.

Solution:

TABLE 3.17 LESS THAN CUMULATIVE FREQUENCY DISTRIBUTION

Scores	No. of Candidates	Score less than	Cumulative Frequency (c.f.)
500 - 550	35	550	35 + 0 = 35
550 – 600	27	600	35 + 27 = 62
600 – 650	24	650	62 + 24 = 86
650 – 700	17	700	86 + 17 = 103
700 – 750	28	750	103 + 28 = 131
750 – 800	19	800	131 + 19 =150

TABLE 3.18 MORE THAN CUMULATIVE FREQUENCY DISTRIBUTION

Scores	No. of Candidates	Score More than	Cumulative Frequency (c.f.)
500 - 550	35	500	150 - 0 = 150
550 – 600	27	550	150 – 35 =115
600 – 650	24	600	115 – 27 = 88
650 – 700	17	650	88 – 24 = 64
700 – 750	28	700	64 – 17 = 47
750 – 800	19	750	47 – 28 = 19

Example 3.7: Calculate the cumulative frequency for the following frequency distribution:

Marks Obtained	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students	5	9	13	7	11

Solution:

TABLE 3.19: CUMULATIVE FREQUENCY DISTRIBUTION

Marks Obtained	No. of Students	Cumulative Frequency (c.f.)
0 – 10	5	5
10 – 20	9	5 + 9 = 14
20 – 30	13	14 + 13 = 27
30 – 40	7	27 + 7 = 34
40 – 50	11	34 + 11 = 45

RELATIVE FREQUENCY DISTRIBUTION

The ratio of the number of times a value occurs in a data set to the total number of outcomes is called relative frequency. The relative frequency can be calculated as per following formula:

$$\textbf{Relative Frequency} = \frac{\textbf{Frequency of a Value/Class}}{\textbf{Total number of Observation}}$$

Example 3.8: Calculate the relative frequency for the following frequency distribution:

Marks Obtained	15 – 25	25 – 35	35 – 45	45 – 55	55 – 65
No. of students	7	11	15	9	13

Solution:

TABLE 3.20: RELATIVE FREQUENCY DISTRIBUTION

Marks Obtained	**No. of Students**	**Relative Frequency**
15 – 25	7	7/55*100 = 12.72%
25 – 35	11	11/55*100 = 20%
35 – 45	15	15/55*100 = 27.27%
45 – 55	9	9/55*100 = 16.36%
55 – 65	13	13/55*100 = 23.63%
Total	**55**	**100%**

Example 3.9: Suppose, you gather a simple random sample of 400 households of Delhi city for the record of number of pet dogs in each household. The table given below shows the result as follow:

No. of Pet dogs	1	2	3	4	5
Frequency	150	90	110	30	20

Prepare a relative frequency distribution of given data.

Solution:

TABLE 3.21: RELATIVE FREQUENCY DISTRIBUTION

No. of Pet Dogs	**No. of Students**	**Relative Frequency**
1	150	150/400*100 = 37.5%

2	90	90/400*100 = 22.5%
3	110	110/400*100 = 27.5%
4	30	30/400*100 = 7.5%
5	20	20/400*100 = 5%
Total	**400**	**100%**

BIVARIATE FREQUENCY DISTRIBUTION

The frequency distribution of two variables is known as Bivariate or Two way frequency distribution. For example, when there are m n cells in a two-way table, it means, it has m classes for X variable and n classes for Y variable. In this type of table, the classes of one variable are arranged horizontally and the classes of second variable are arranged vertically. The frequency of each cell is found by paring values both variables (X and Y). Thus, this entire set of cell frequencies forms a bivariate frequency distribution table. Now, let's take an example of bivariate frequency distribution.

Example 3.10: The following data relates to sales and advertisement expenditure of 15 companies. Form a bivariate frequency distribution having class interval 62 – 64, 64 – 66 and so on and 115 – 125, 125 – 135 and so on.

Company	Sales	Adv. Exp.	Company	Sales	Adv. Exp.
1	160	71	9	116	71
2	135	65	10	129	62
3	136	65	11	163	66
4	145	64	12	134	67
5	148	69	13	122	64
6	124	71	14	134	68
7	126	71	15	140	67
8	128	70			

Solution:

To solve the above question, advertisement expenditure (Adv. Exp.) is to be divided into 5 classes and sales is divided into 6 classes. For tabulation of information in appropriate cells, first, row to which advertisement expenditure (X) belong is determined. Afterwards on consideration of sales (Y), the column in which it should be included is determined. Thus, the bivariate frequency table can be prepared as follows:

TABLE 3.22: TWO WAY FREQUENCY SHOWING SALES AND ADVERTISEMENT EXPENDITURE OF 15 COMPANIES

Sales (X) / Adv. Exp. (Y)	115 -125	125 – 135	135 – 145	145 - 155	155 - 165	Total (f)
62 – 64	---	I (1)	---	---	---	1
64 – 66	I (1)	---	II (2)	I (1)	---	4
66 – 68	---	I (1)	I (1)	---	I (1)	3
68 – 70	---	I (1)	---	I (1)	---	2
70 – 72	II (2)	II (2)	---	---	I (1)	5
Total (f)	3	5	3	2	2	15

Marginal distribution of X and Y: The frequency distribution of the values of the sales (X) together with their frequency totals (f_X) and the frequency distribution of the values of the advertisement expenditure (Y) together with their frequency totals (f_Y) is known as the marginal frequency distribution of variable X and Y respectively.

TABLE 3.23: MARGINAL DISTRIBUTION OF X AND Y

Marginal distribution of Y		Marginal distribution of X	
Adv. Exp. (Y)	**f**	**Sales (X)**	**f**
62 - 64	1	115 – 125	3
64 - 66	4	125 – 135	5
66 - 68	3	135 – 145	3
68 - 70	2	145 – 155	2
70 - 72	5	155 – 165	2
Total	**15**	**Total**	**15**

Conditional Distribution of X and Y: To calculate the conditional frequency distribution of *X* for a given value of *Y*, *we take* the values of *X* together with their frequencies corresponding to the fixed values of *Y*. Alike, we can obtain the conditional frequency distribution of *Y* for given values of *X*. Table 3.24 shows conditional distribution of variable X and Y of the example 3.10 given above.

TABLE 3.24: CONDITIONAL DISTRIBUTION OF X AND Y

Conditional distribution of X when Y = 66 - 68		Conditional distribution of Y when X = 145 - 155	
Adv. Exp. (X)	**f_X**	**Sales (Y)**	**f_Y**
62 - 64	0	115 – 125	0
64 - 66	1	125 – 135	1
66 - 68	1	135 – 145	0
68 - 70	0	145 – 155	1
70 - 72	1	155 – 165	0
Total	**3**	**Total**	**2**

3.4 CONCEPT OF TABULATION OF DATA

It is a creative method of presenting the data in a summarized and understandable form. It provides full information in the least possible space without losing the authenticity of the data. As a final stage of collection of the data, tabulation is useful for further statistical analysis and interpretations. In a table, rows are the horizontal arrangement whereas the columns are the vertical arrangement of data. The process of tabulation can be simple or complex according to categorization. Some of the important definitions of tabulation are given below:

According to 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."

In the words of **Professor Bowely,** "It is intermediate process between the collection of data in whatever form they are obtained, and the final reasoned account of the result shown by the statistics."

3.4.0 OBJECTIVES OF TABUALTION

The main objectives of tabulation are:

- **Simplify the complexity of the data:** The statistical data in its original form is complex and meaningless with unnecessary details. To simplify the complex data, tabulation procedure is used to systematically present the data in the form of columns and rows.
- **Facilitate comparative analysis of data:** Tabulation process is helpful in comparative analysis. It is highly necessary to make a comparative study of the data to extract useful hidden information by representing the relevant figures, their sub-totals, and totals in the form of columns and rows. This is very much facilitated by the technique of tabulation.
- **Ensure economy of space and time:** The statistical data in its original form is always described through notes and phrases which take a lot of space and time. It is important to avoid needless details and repetitions of data to economize the space and time. The tabulation process is helpful in saving of pace and time without sacrificing the quality and utility of the data.
- **Indicate the trend and pattern of data:** With the help of tabulation process, the trend and pattern of the data relating to the phenomenon is easily determined. Tabulation is helpful in identification of trend and pattern in the data.
- **Facilitate references:** There are two types of sources of data – Primary and secondary. If in any study secondary data is used, it is essential to specify the source from which the same data have been obtained from a table which is properly numbered and market with a title. Tabulation helpful in helpful in making of such type of references.
- **Facilitate computation of various factors:** For a proper statistical analysis and interpretation, computation of various factors like, average, dispersion, skewness kurtosis, and correlation is required. The tabulation of the data is very helpful in computation of all such factors.

- **Detect Errors:** During data collection process, lots of errors occurred. These errors cannot be easily detected. To detect the errors, tabulation is done and also helpful in testing the accuracy of the data.

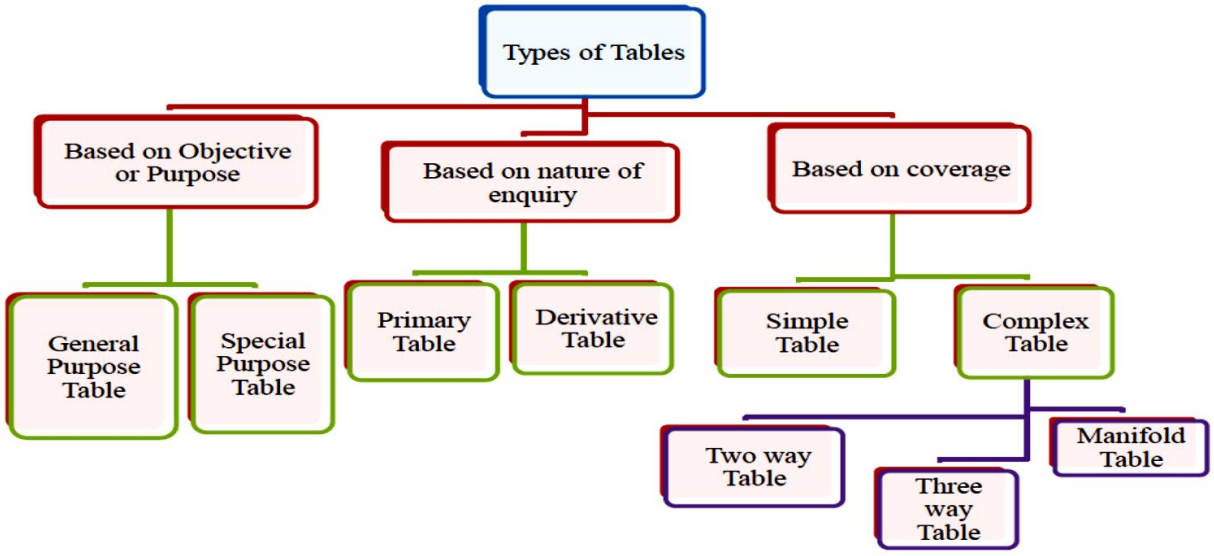
3.4.1 TYPES OF TABULATION

In statistics, various types of tables are made. The choice of tables mainly depends on the following three bases:

- Based on objective or purpose
- Based on nature of the enquiry
- Based on coverage (Simple and Complex)

Figure 3.3 portrays the various bases of types of tabulation:

FIGURE 3.3: TYPES OF TABLES



Based on Objective or Purpose:

Following two types of tables are made herein:

- **General Purpose Table:** It is also known as reference or informative table. It is an appropriate mode of collecting and presenting a systematically arranged and chronologically ordered data suitable for ready reference and record. These tables are of repository nature and mainly used by researchers, statisticians and so on.

- **Special Purpose Table:** Also known as summary or interpretative table is analytical in nature. It is prepared for comparative studies. It is helpful in studying the relationship between figures or data for a specific purpose. In such tables, interpretative measures like ratios, percentages and so on are used for comparisons.

Based on Nature of Enquiry:

It can be of two types:

- **Primary Table:** It is also known as original table. In a primary table, first hand data in its original form is recorded. Instead of ratios or percentages, the actual and absolute facts are recorded in a primary table. For example, the data collected in a population survey may be expressed in a primary table.
- **Derivative Table:** It recorded the figures and facts in terms of ratio, percentage, aggregates from the primary table. In derivative table, first-hand information is used in modified form for other purpose. For example, the table showing trend value, seasonal and cyclical variations is a derivative table.

Based on Coverage:

On the basis of this, the table is divided into two parts:

- **Simple Table:** It is also known as one way table. In simple table, the data is classified on the basis of a single characteristic of phenomenon under study.
- **Complex Table:** In complex table, the data is classified on the basis of two or more than two characteristics of phenomenon under study. It can be of three types.

Two way Table: In two way table, the data is classified on the basis of only two characteristics of phenomenon under study.

Three way Table: In three way table, the data is classified on the basis of two or more characteristics of phenomenon under study.

Manifold Table: In manifold table or high order table, the data is categorized as per many characteristics of phenomenon under study.

3.4.2 COMPONENTS OF A TABLE

For the construction of a table, it is important to have knowledge about the all components of table. These components put together form a table. To construct a table, the data is arranged in rows and columns with some clarifying notes. Table 3.24 shows some of the important components of a good table:

TABLE NO 3.24: TABLE FORMAT

TITLE

(Head note or prefatory note, if any)

Stub Heading	Caption					Total
	Sub Head		Sub Head			
	Column Head	Column Head	Column Head	Column Head	Column Head	
Stub Entries	Main Body					
Total						

Foot Note:
Source Note:

- **Table Number:** It is used for identification of a table. If there are more than one table in any study, table number is helpful in differentiates one table from another. It is mentioned at top or in beginning with title of the table.
- **Title:** The title describes about the contents of the table. It is shown next to the table number mentioned either at the top of the table or below it. It should be brief, vibrant and carefully formulated to make unbiased and unambiguity interpretations from the table. It should be precisely described nature, place, time and sources of data.
- **Head Notes or Prefatory Notes:** As per requirement, head note also known by prefatory notes. It is given, just below the title of the table to further describe the contents in detail. Headnotes are usually centrally aligned and enclosed in brackets.
- **Captions or Column Headings:** Captions are the heading or sub heading used for vertical columns which explain what the column represents. It should be brief, concise and self-explanatory. It usually written in the middle of the columns in small letters. It may be used as a head note along with the title or may also be indicated in the columns or rows of the table.

- **Stubs or Row Headings:** Stubs are the heading or sub heading used for horizontal rows which tells about what the rows represent. It kept narrow as possible without losing accuracy and clearness of the data.
- **Main Body of the Table:** It is a key component of a table. It represents numerical information in the form of rows and columns. It should not include undesirable and irrelevant fact to increase the usefulness of table.
- **Unit of Measurement:** It is stated along with the title if same unit is used for entire table. If different unit is given for row or column of the table, it must be stated along with 'stubs' or 'captions'. If figures are large, they must be rounded off and the method of rounding should be clearly indicated.
- **Foot Note:** If something is not understandable to the reader from the title, captions and stubs, it should be explained in footnotes. It is the main purpose of foot notes. It is placed at the bottom directly below the body of the table. For the identification of footnotes, various types of systems are used. One of them is use of small numbers like 1, 2, 3, or letters like a, b, c, d and so on. You can also use one star (*) for first foot note, two star (**) for second footnotes, three star (***) for third footnotes and so on.
- **Source Note:** It is a brief statement which indicates the source of data used in the table. It is mentioned at the bottom below the footnote. The main purpose of source note is to satisfy the secondary data users about the accuracy and reliability of the figures.

3.4.3 ESSENTIALS OF A GOOD TABLE

Prof. Bowley has rightly pointed out "In collection and tabulation, common sense is the chief requisite and experience is the chief teacher." Following points should be kept in mind for making a table:

1. It should include all the essential components of a table like table number, title, body, and source note etc.,.
2. It should be compact, complete, self-explanatory and simple to understand.
3. It should be concised and not burdened with unnecessary details.
4. Rows and columns of a table should be numbered.
5. The captions and stubs should be arranged alphabetically or chronologically.
6. The unit of measurement should be mentioned in the head note.

7. The figures should be rounded off to the nearest hundred, or thousand or lakh.

8. In case of non-availability of information, one should write N.A. or indicate it by dash (-).

9. The expression 'etc' should be avoided in a table.

3.5 CHECK YOUR PROGRESS

Fill in the Blanks

1. Variables are of two kinds discrete and ---------**.**

2. The class mid-point is the value laying half way between the upper and --------- of the class.

3. Table number is given to a table for ---------- purpose.

4. ---------- means ordering raw data into groups or classes as per their similarities.

5. The data can be classified into inclusive and ---------- types.

3.6 SUMMARY

The raw data is known as unmodified form of data. For the further statistical analysis, raw data should be modified with the help of classification, tabulation and graphication. These are the techniques used for the presentation of raw data in a simple and summarized form. It should be easily understandable to a layman.

The systematic arrangement of raw data on the basis of similarities is known as classification of data. Condensation of data, comparison facilitation, study of relationship between two or more variables and use for statistical treatment are the main functions of classification. For effective implementation of classification technique, various rules are followed. According to these rules, data must be unambiguous, exhaustive and mutually exclusive in nature. It should be suitable to the purpose of enquiry and its stability and flexibility should also be maintained. On the basis of nature and purpose of statistical enquiry, the raw data can be classified geographically, chronologically, quantitatively and qualitatively.

The raw data is also organized in the form of frequency distribution. It is a comprehensive way of presenting raw data of quantitative (Discrete or Continuous) variable. Various classes, class frequencies,

class interval, class limit and class mid-point is the important terms used in frequency distribution. Frequency distribution can be of many types. The ungrouped, grouped, cumulative, relative and bivariate are some of them.

For increasing the usefulness and quality of data within minimum space and efforts, tabulation technique is used. It is a method of arranging raw data in the form of rows and columns. Simplification, comparison, economy of space and time, predict trend and pattern, computation, finding errors and facilitating references are main objectives of tabulation. The tables are categorized on the basis of objective, nature and coverage of enquiry. According to objective or purpose, general and special purpose tables are made. Primary and Derivative tables are made on the basis of nature of enquiry. Similarly, as per coverage, simple and complex tables are made. Further, complex table can be formed in three ways: Two way table, three way table and manifold table. Table number, title, head notes, captions, stubs, main body of content, unit of measurement, foot note and source notes are the important components of a table. The table should not be overloaded with too much detail. Overall, a table should be easily understandable, compact and self-explanatory.

3.7 KEYWORDS

- **Class Boundary:** It is upper and lower limit of a class interval.
- **Class Frequency:** It refers to the number of times a value is frequently occurring in a data set.
- **Data Array:** It is arrangement of raw data either in ascending or descending order.
- **Raw Data:** It refers to unanalyzed and unprocessed data by statistical methods.
- **Tabulation:** It is a method of presenting the data in the form of rows and columns in a summarized way.
- **Class width:** It is calculated by dividing the difference of upper limit and lower limit of a class divided with the number of class interval.

3.8 SELF ASSESSMENT TEST

1. Explain the role of tabulation in presenting the data.

2. What are the characteristics of a good table?

3. Define classification. State its important objectives.

4. Describe different types of classification with suitable examples.

5. What is tabulation? What are the different bases of tabulation?

6. Briefly explain the essential parts of a table. What are the basic rules of making a statistical tables?

7. What is the grouped and ungrouped frequency distribution? Explain with a suitable example.

8. What is inclusive series? How can you convert an inclusive series into an exclusive series? Give a suitable example.

9. What is frequency distribution? What are the various types of frequency distribution?

10. A market survey was conducted of 40 respondents on acceptability of a new product. The following scores on an appropriate scale recorded:

9	54	14	54	38	93	18	34	52	4
2	17	32	42	31	92	48	44	14	4
3	53	31	52	53	44	24	34	49	3
3	23	32	72	52	62	33	73	41	3

Prepare a frequency distribution table.

11. The marks obtained by 30 students in statistics test are given below:

4	34	75	33	45	36	31	25	95	4
6	76	92	26	27	26	68	35	59	7
1	52	43	54	57	74	34	23	62	9

Prepare a frequency table and cumulative frequency table.

3.9 ANSWER TO CHECK YOUR PROGRESS

Answer to Fill in the Blanks

1. Continuous

2. Lower limits

3. Identification

4. Classification

5. Exclusive

3.10 REFERENCES/SUGGESTED READINGS

- Black, K. (2023). *Business statistics: for contemporary decision making*. John Wiley & Sons.
- Bluman, A. (2014). *Elementary Statistics: A step by step approach 9e*. McGraw Hill.
- Groebner, D. F., Shannon, P. W., Fry, P. C., & Smith, K. D. (2008). *Business statistics*. Pearson education.
- Gupta, A. (2021). *Business statistics*. RAJEEV BANSAL.
- Bajpai, N. (2009). *Business statistics*. Pearson Education India.
- Sharma, J. K. (2012). *Business statistics*. Pearson Education India.

Subject: Business Statistics-1	
Lesson No: 4	**Author: Dr. Vizender Singh**
Subject Code: BCOM 302	**Vetter: Prof. Kuldeep Bansal**

PRESENTATION OF DATA

STRUCTURE

4.0 LEARNING OBJECTIVES

After reading this chapter you will be able to understand:

- Presentation of data using diagrams, their benefits and limitation
- Different types of diagrams
- Presentation of data using graphs, their benefits and limitation
- Different type of graphs

4.1 INTRODUCTION

In the previous chapter we have seen how to condense the mass data by the method of classification and tabulation. It is not so easy to understand figures every time and these may not be interesting for everyone. Further, some of figures are very confusing and complicated which creates problems while analyzing. One of the most convincing and appealing ways in which statistical results may be represented is through graphs and diagrams. Diagrams can be more easily compared, and can be interpreted by a layman. Diagrams are more attractive and have a visual appeal. This is the reason that diagrams are often used by businessmen, newspapers, magazines, journals, government agencies and also for advertising and educating people. In this chapter we will discuss about the diagrammatic presentation of data and graphical presentation of data.

William play fair - 1759 - 1823 said in his 'Commercial and political Atlas of 1801' that "Information take days through data presented in tables, but can be obtained in minutes through diagrams."

Cold figures are uninspiring to most people. Diagrams help us to see the pattern and shape of any complex idea, just as a map gives a bird's eye view of wide sketch of a country, so diagrams help us to visualize the whole meaning of a numerical complex at single glance. Diagrams register a meaningful impression almost before you think ....." (Moroney).

"The important point that must be borne in mind at all times that the pictorial representation chosen for any situation must depict the true relationship and point out the proper conclusion. Above all the chart must be honest." .... C. W. LOWE.

Data can be visually presented as shown in the Fig.1, i.e. Diagrammatic and Graphic. These are also known as visualization because they create an image in the mind of the reader. Both of these representation of data further categorized among different parts. This classification is discussed in the next section of this chapter.

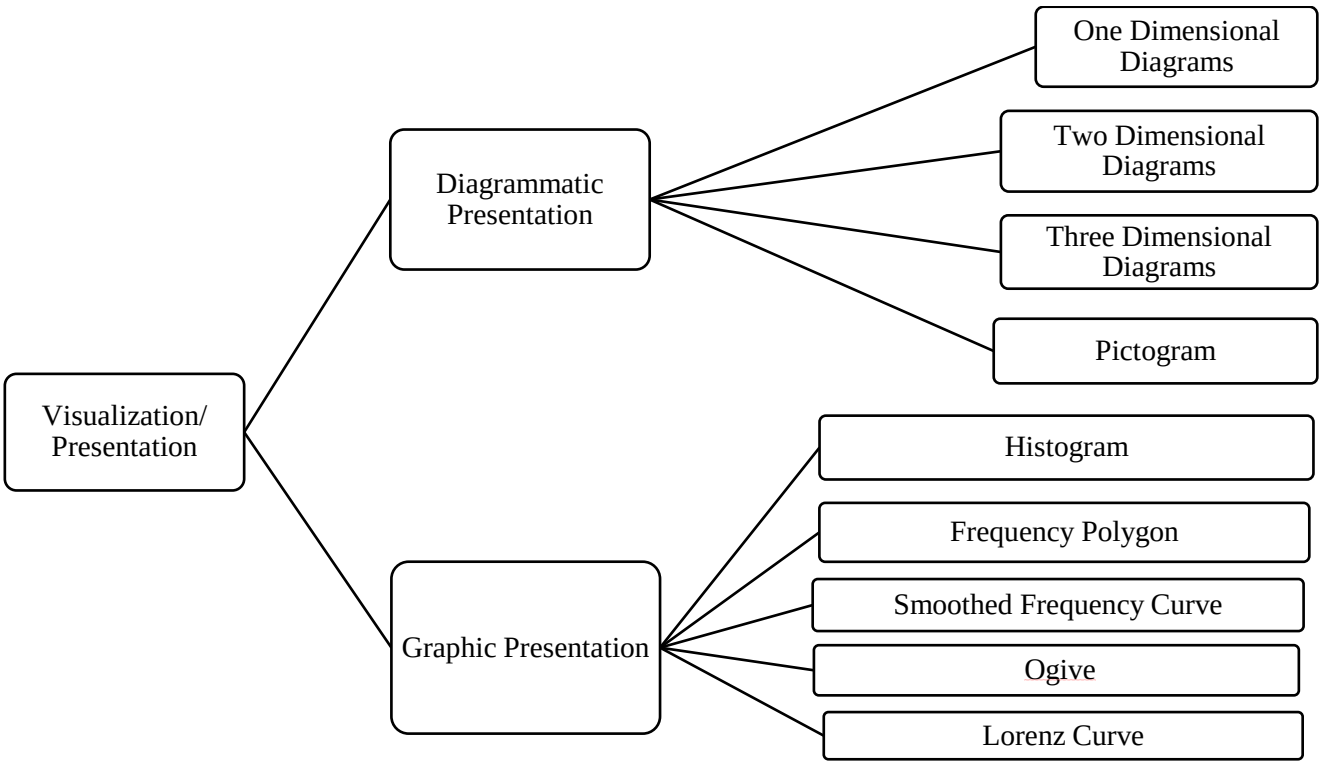


Fig.1 Types of Data Presentation/Visualization

These are discussed one by one in this chapter. First of all we will discuss about Diagrammatic Presentation of data and later on we will focus on Graphical Presentation of data.

4.2 DIAGRAMMATIC PRESENTATION OF DATA

Diagrammatic Presentation of Data provides an immediate understanding of the real situation. Diagrammatic presentation of data converts the highly complex ideas included in numbers into more concrete and quickly understandable form. Diagrams may be less certain but are much more efficient than tables in displaying the data. There are many kinds of diagrams in general use such as Pictograms, Cartograms, Bar Diagrams & Pie Diagrams etc. Diagrams help in visual comparison and have a bird's eye view. Diagrams are different geometrical shape such as bars, circles etc. Diagrams are based on scale but are not confined to points or lines. They are more attractive and easier to understand than tables and graphs.

4.2.1 BENEFITS OF DIAGRAMMATIC PRESENTATION OF DATA

1. Diagrammatic presentation is very attractive and easy to understand.
2. There is no need of technical knowledge and expertise to form and understand diagrams.
3. Diagrams require less time and efforts as compared to graphs.
4. Visual things are more memorisable and had a lasting impression.
5. Diagrams can be understood by everyone because there is no language barrier.
6. Diagrams are helpful to represent huge data in a simplified and intelligible form.
7. Diagrams are helpful in making comparison of data.
8. Diagrams also provides hidden information about the data.

4.2.2 GENERAL RULES FOR CONSTRUCTING DIAGRAMS

1. The diagrams should be simple and clear.
2. There should be a suitable title without damaging clarity.
3. You should keep in mind the height and width must be maintained forming a diagram.
4. You should select a proper scale such as it may be in even numbers or multiples of five or ten.
5. There should be footnotes under the diagram.
6. An index should be inserted for explaining different lines, shades and colours.

4.2.3 TYPES OF DIAGRAMS

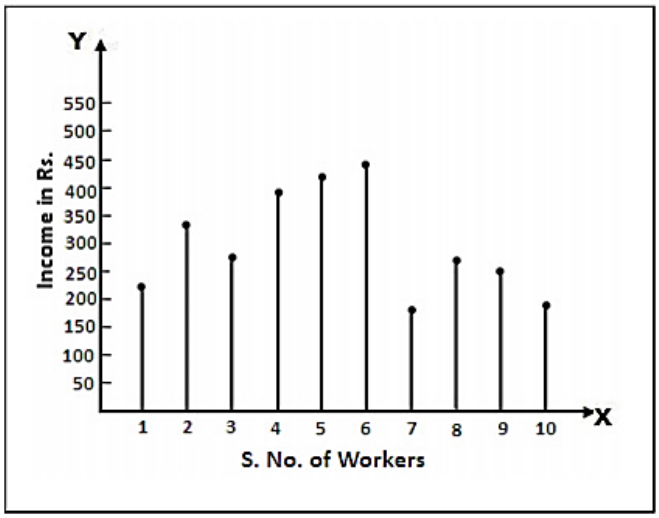
As we have discussed earlier in the figure 1, diagrammatic presentation can be divided into four parts which are explained below:

I. One Dimensional Diagrams

One dimensional diagrams are the diagrams in which only one dimension is included i.e. height. Here, width or thickness is not measured. These diagrams may be seen in the form of lines or bars. These diagrams are divides into five parts discussed as under:

Line Diagram refers to the diagram where you have to present many items and there is not much difference in their values. These are simply formed through using a vertical or horizontal line for each item according to scale and the space among lines is kept uniform. This is the easiest method of data representation which makes comparison easy but it is not an attractive method. It can be shown as diagram I:

Here, Income of 10 workers is shown according to their series number. Income of worker is shown on Y axis and Series number of workers is shown on X axis.

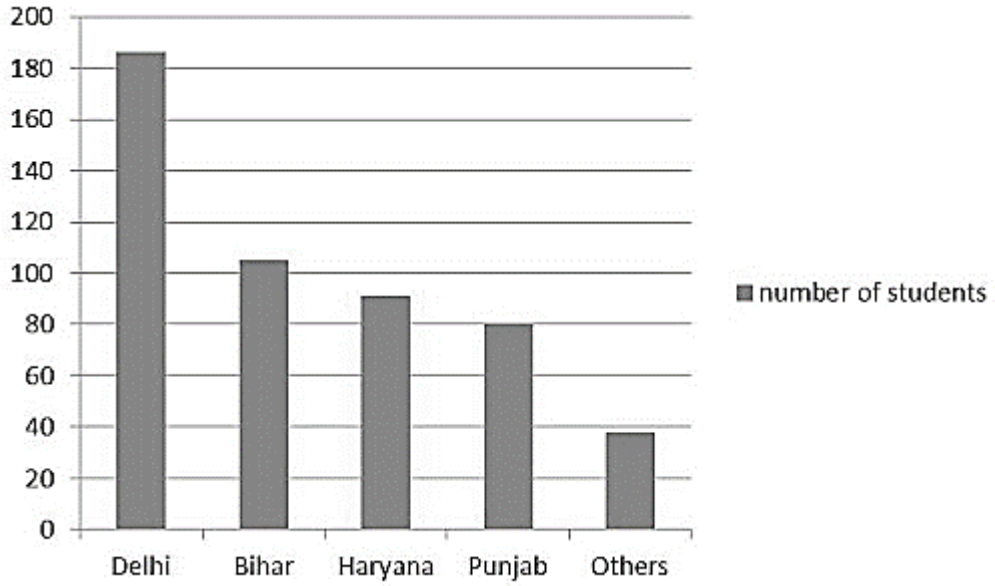


Line Diagram of Income of 10 workers

Diagram I

Simple Bar Diagram represents only one variable. For example sales, production, population figures etc. for various years may be shown by simple bar charts. Since these are of the same width and vary only in heights or lengths. Further, it becomes very easy for readers to study the relationship. Simple bar diagrams are very popular in practice. A bar chart can be either vertical or horizontal.

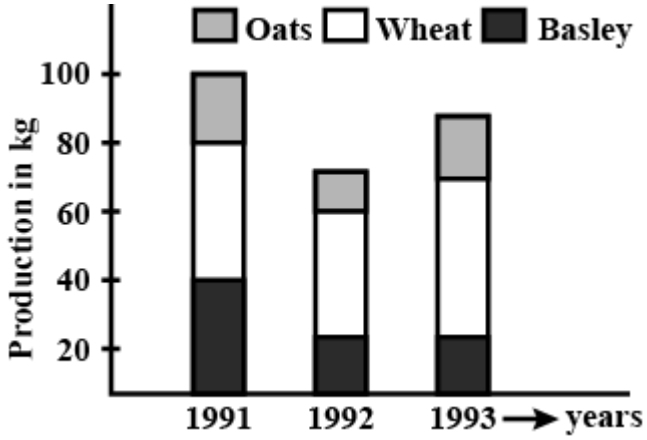
For example, Diagrammatic representation of data relating to number of students in different states is given below:



Simple Bar Diagram II

Sub - divided Bar Diagram: While constructing such a diagram, the various components in each bar should be kept in the same order. A common and helpful arrangement is that of presenting each bar in the order of magnitude with the largest component at the bottom and the smallest at the top. The components are shown with different shades or colors with a proper index.

For example, Sub-divided bar diagram is shown production of Oats, Wheat and Basley in different years.



Sub-divided Bar Diagram III

Multiple Bar Diagram: This method can be used for data which is made up of two or more

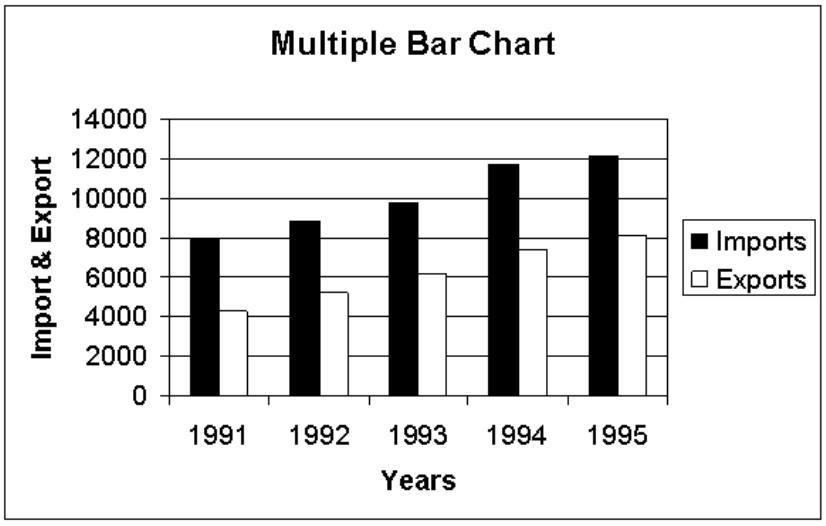
components. In this method the components are shown as separate adjoining bars. The height of each bar represents the actual value of the component. The components are shown by different shades or colors. Where changes in actual values of component figures only are required, multiple bar charts are used.

For example, Draw a multiple bar chart to represent the imports and exports of Canada (values in $) for the years 1991 to 1995.

Years	Imports	Exports
1991	7930	4260
1992	8850	5225
1993	9780	6150
1994	11720	7340
1995	12150	8145

Table I: Import and Export data of Canada from 1991 to 1995

Sol:

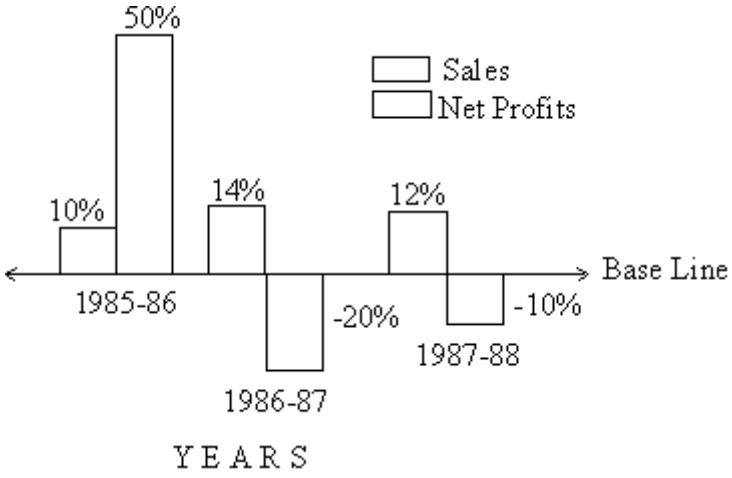


Multiple Bar Diagram IV

Deviation Bar Diagram: Deviation bars are used to represent net quantities - excess or

deficit i.e. net profit, net loss, net exports or imports, swings in voting etc. Such bars have both positive and negative values. Positive values lie above the base line and negative values lie below it.

For example, Sales and Net Profit are shown in the Deviation Bar Diagram.



Deviation Bar Diagram V

All the above diagram represents one dimensional Diagrams. In the next part, we will discuss two dimensional Diagrams.

II. Two Dimensional Diagrams

Two dimensional diagrams are those diagrams where both the dimension length as well as width of the bar are considered for construction of diagrams. These diagrams are also known as "Area" or "Surface" diagrams. There are three types of area diagrams such as Rectangles, Squares and Pie Diagrams.

Rectangles are diagrams which are used to represent the magnitude of two or more values and rectangles are placed side by side so that comparison can be made. These diagrams represents two different characteristics of data. We may represent data as they are given or these can be converted into percent and then subdivided into parts according to the length.

For Example, Represent the following data by sub-divided percentage rectangular diagram.

Items of Expenditure	Family A (Income Rs.5000)	Family B (income Rs.8000)
Food	2000	2500
Clothing	1000	2000
House Rent	800	1000
Fuel and lighting	400	500
Miscellaneous	800	2000
Total	5000	8000

Table II: Items of expenditure with Income of Family A & B

Sol: First of all, these data are converted into percentage and then subdivided into different section of rectangular diagram.

Items of Expenditure	Family A		Family B	
	Rs.	Y	Rs.	Y
Food	2000	40	2500	31
Clothing	1000	20	2000	25
House Rent	800	16	1000	13
Fuel and Lighting	400	8	500	6
Miscellaneous	800	16	2000	25
Total	5000	100	8000	100

Table III: Items of expenditure with Income of Family A & B

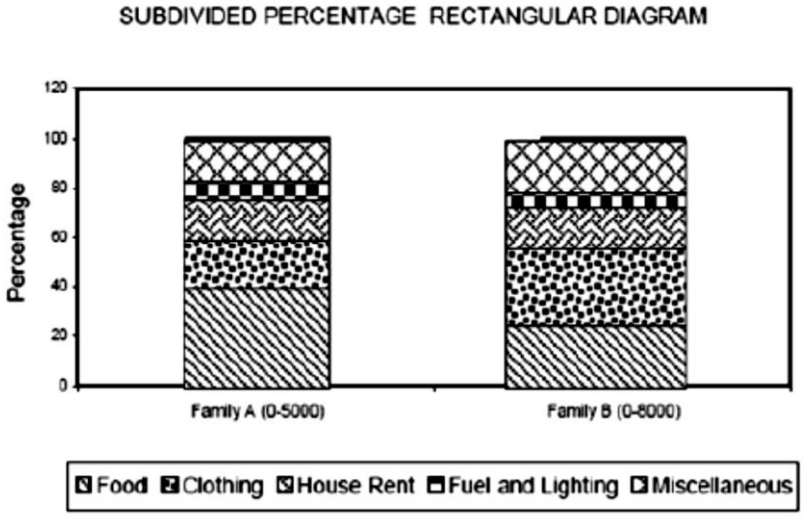


Diagram VI: Subdivided Percentage Rectangular Diagram

Square diagrams requires square root of the given data which provides the measurement of the sides of the square. For example, Yield of rice in Kgs. per acre of five countries are:

Country	U.S.A	Australia	U.K	Canada	India
Yield of rice in Kgs per acre	6400	1600	2500	3600	4900

Table IV: Yield of Rice in Kgs Per Acre of Different Countries

Represent the above data by a square diagram.
Sol: First of all we calculate the square root of data as follows:

Country	Yield	Square root	Side of the square in cm
U.S.A	6400	80	4
Australia	1600	40	2
U.K.	2500	50	2.5
Canada	3600	60	3
India	4900	70	3.5

Table V: Yield of Rice in Kgs Per Acre of Different Countries

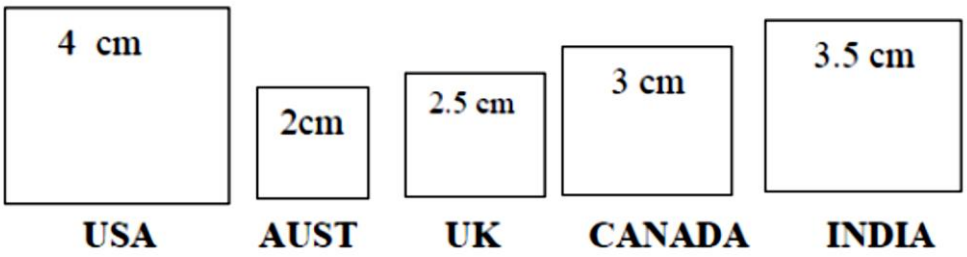


Diagram VII: Square Diagram Representing Yield of Rice in Kgs Per Acre of Different Countries

Pie Charts or Diagram consists of a circle in which the radii divide the area into sectors. Further, these sectors are proportional to the values of the component items under investigation. Also, the whole circle represents the entire data under investigation.

Steps to draw a Pie Chart

- Express the different components of the given data in percentages of the whole
- Multiply each percentage component with 3.6 (since the total angle of a circle at the center is 360°)
- Draw a circle
- Divide the circle into different sectors with the central angles of each component
- Shade each sector differently

Let us take **an example**, consider the yearly expenditure of a Mr. Ted, a college undergraduate.

Tuition fees	$ 6000
Books and lab.	$ 2000
Clothes / cleaning	$ 2000
Room and boarding	$ 12000
Transportation	$ 3000
Insurance	$ 1000
Sundry expenses	$ 4000
Total expenditure	$ 30000

Table VI: Yearly Expenditure

Now as explained above, we calculate the angles corresponding to various items (components).

Tuition fees = $\frac{6000}{30000} \times 360^0 = 72^0$

Book and lab = $\frac{2000}{30000} \times 360^0 = 24^0$

Clothes / cleaning = $\frac{2000}{30000} \times 360^0 = 24^0$

Room and boarding = $\frac{12000}{30000} \times 360^0 = 144^0$

Transportation = $\frac{3000}{30000} \times 360^0 = 36^0$

Insurance = $\frac{1000}{30000} \times 360^0 = 12^0$

Sundry expenses = $\frac{4000}{30000} \times 360^0 = 48^0$

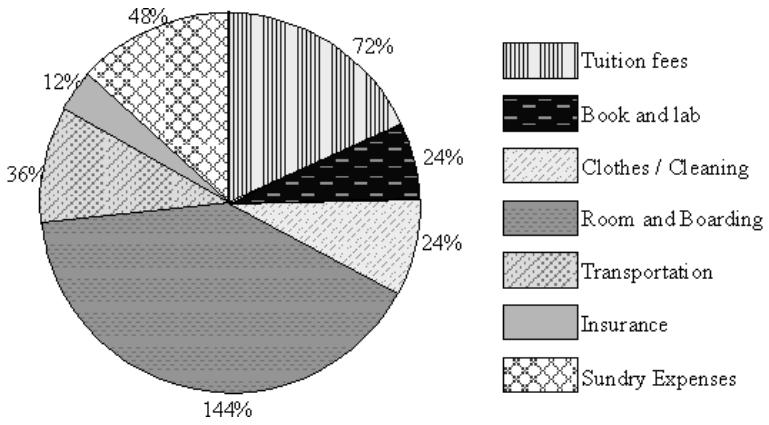


Diagram VIII : Pie Chart of total expenditure

III. Three Dimensional Diagrams

Three dimensional diagrams are the diagrams in which three dimension are taken into account. These dimensions are length, width and height. These diagrams are also known as Cubic Diagram and these diagrams may be drawn in the form of cylinders, blocks, spheres, etc.

IV. Pictogram

As the name suggested that pictogram uses appropriate pictures to represent data. It is also known as Picture Graph or Pictograph. These are very attractive and create a lasting effect on viewer mind.

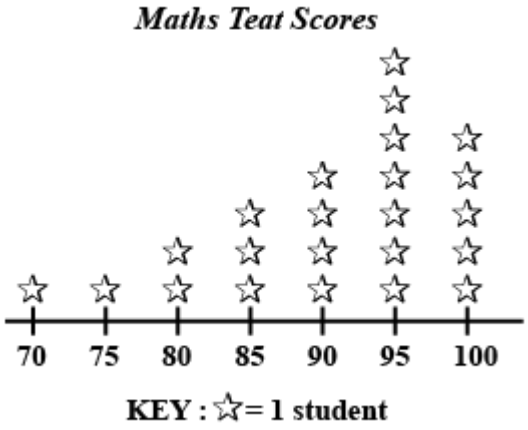


Diagram IX : Pictogram representing Math Test Score

4.2.4 LIMITATIONS OF DIAGRAMMATIC PRESENTATION OF DATA

As we studied that the above diagrams are very attractive as well as very easy to form. But, still there are some deficiencies which are explained below:

1. These diagrams do not provide detailed information.

2. Diagrams can be easily misinterpreted.

3. Diagrams can take much time and labour.

4. Exact measurement is not possible in diagrams.

4.3 GRAPHIC PRESENTATION OF DATA

A graph is a visual representation of data by a continuous curve on a graph paper. Graphs are more attractive than a table or figure and revealing their inner pattern. A common man can even understand the message given in the graph. Graphical Representation is a way of analysing numerical data. It exhibits the relation between data, ideas, information and concepts in a diagram. Graphs enable us in studying the cause and effect relationship between two variables. Graphs help to measure the extent of change in one variable when another variable changes by a certain amount. It is easy to understand and it is one of the most important learning strategies. It always depends on the type of information in a

particular domain. There are different types of graphical representation. There are some important forms of graphs i.e. Histogram, Frequency Polygon, Smooth Frequency Curve, Ogive and Lorenz Curve. These are discussed in detail in the next section i.e. Type of Graphs.

4.3.1 BENEFITS OF GRAPHIC PRESENTATION OF DATA

1. Graphs represent complex data in a simple form.
2. Values of median, mode can be found through graphs.
3. Graphs create long lasting effect on people's mind.
4. Graphs are attractive and impressive.
5. Graphs make data simple and intelligible.
6. Graphs make comparison possible.
7. Graphs save time and labour.
8. Graphs have universal utility.
9. Graphs give more information.

4.3.2 GENERAL RULES FOR CONSTRUCTING A GRAPH

1. First of all, one should focus on the title of a graph which should be simple and clear.
2. Then it is important to mention the measurement unit in the graph so that values can determined easily.
3. It is necessary to choose a proper scale so that we can present data in an accurate manner.
4. It is important to provide different colours, shades, designs in the graph for better understanding.
5. In the bottom of graph, source of information must be given.
6. Graph should be simple and clear without any ambiguity.

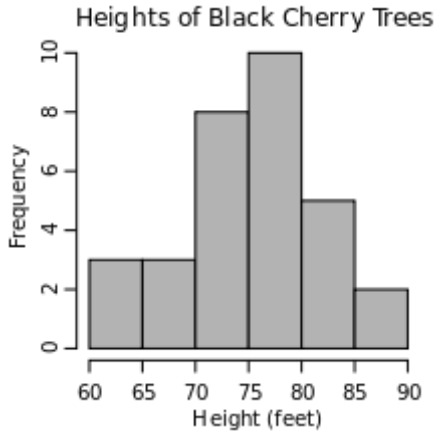
7. Graph is a form of visualisation aid for presentation of data so, one must choose correct size, letters and colours.

4.3.3 TYPES OF GRAPHS

There are major five types of graphs which are explained below with the help of an example:

I. Histogram

A histogram is a graph showing the frequency of occurrence of each value of the variable being analysed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the 'X-axis' and the frequencies on the 'Y-axis'if the classes are of equal width and frequency density (f/c) on 'Y-axis' if the classes are of un equal width. The height of each rectangle represents the frequency or frequency density of the class interval. Each rectangle is formed with the other so as to give a continuous picture. Such a graph is also called staircase or block diagram. However, we cannot construct a histogram for distribution with open-end classes.

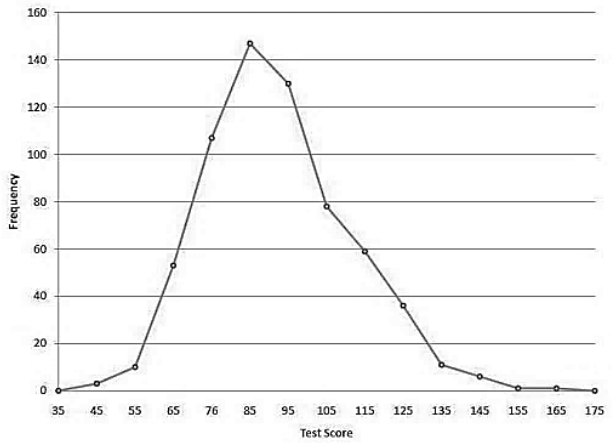


Histogram

II. Frequency Polygon

If we mark the midpoints of the top horizontal sides of the rectangles in a histogram and join them by a straight line, the figure so formed is called a Frequency Polygon. This is done under the assumption that the frequencies in a class interval are evenly distributed throughout the class. The area of the polygon is equal to the area of the histogram, because the area left outside is just equal to the area included in it. Another method of drawing frequency polygon is on the X axis

draw the mid points and on the Y axis the frequency density (f/c) join the points by straight line to obtain frequency polygon.



Frequency Polygon

III. Smoothed Frequency Curve

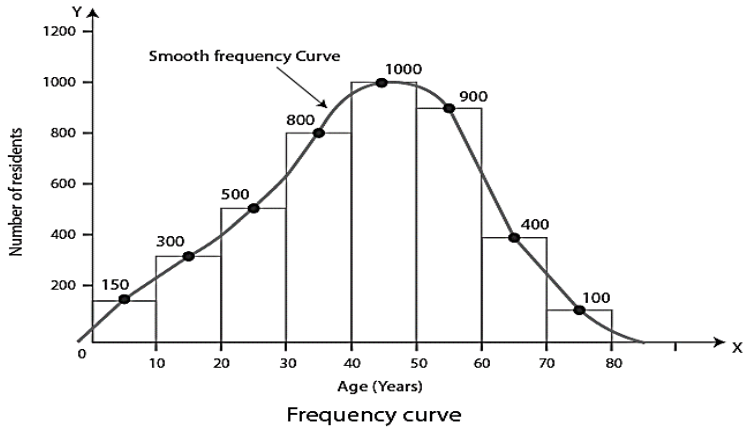
If the middle point of the upper boundaries of the rectangles of a histogram is corrected by a smooth freehand curve, then that diagram is called frequency curve. The curve should begin and end at the base line.

For example,

Make a frequency curve of the following data.

Age (Years)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No of Residents	150	300	500	800	1,000	900	400	100

Sol: The given data set is first converted into a histogram. Mid-points o the top of the rectangles of the histogram are marked. These points are joined through a freehand smoothed curve, as given below.



Smooth Frequency Curve

IV. Ogive

The cumulative frequency gives the cumulative frequency of each of the class. The curve table is obtained by plotting cumulative frequencies is called a cumulative frequency curve or an ogive. There are two type of ogive namely:

1. The 'less than ogive'

   In less than ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve.

2. The 'more than ogive'

   In more than ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

   For example,

Weekly wages (₹)	No of workers
0-20 20-40 40-60 60-80	10 20 40 20

80-100	10

Draw the 'less than' and 'more than' ogive on the same graph paper from the data given below:

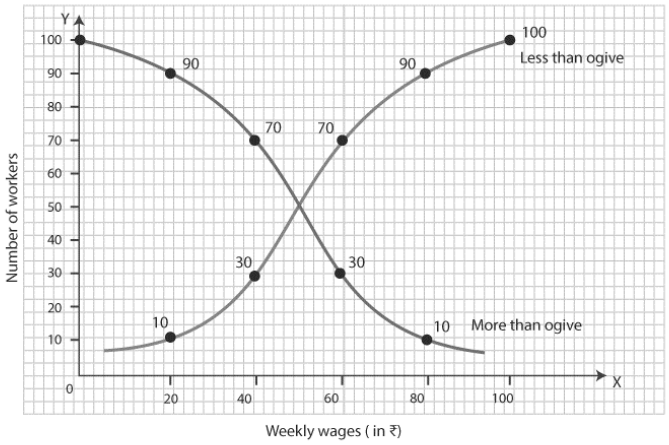
Solution: (i) 'Less than' method

Weekly wages (₹)	C.F
Less 20	10
Less 40	30
Less 60	70
Less 80	90
Less 100	100

(ii) 'More than' method

Weekly wages (₹)	C.F
Less 20	10
Less 40	30
Less 60	70
Less 80	90
Less 100	100

Both 'less than' and 'more than' ogives based on the above data are presented in the following graph.



'Less than' and 'More than' Ogives

V. Lorenz Curve

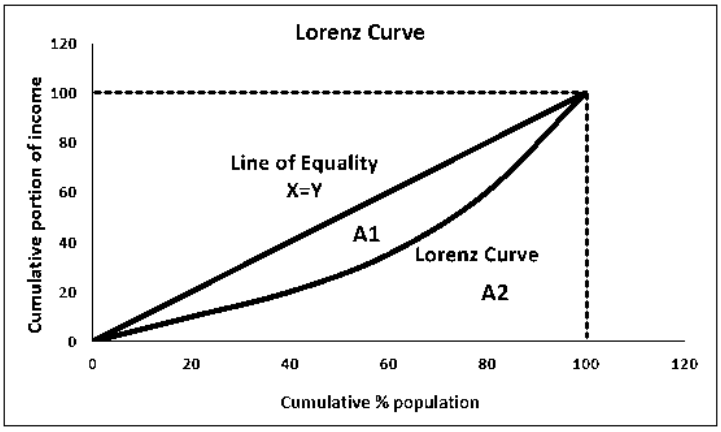
It is a graphical method of studying dispersion among data and this curve was introduced by Max. O. Lorenz, a great Economist as well as Statistician. It is a percentage of cumulative values of one variable in combined with the percentage of cumulative values in other variable and then Lorenz curve is drawn. This curve starts with the origin (0,0) and ends at (100,100). For example, Let us consider an economy with the following population and income statistics:

Population %	Income Portion %
0	0
20	10
40	20
60	35
80	60
100	100

And for the line of perfect equality, let us consider this table:

Population %	Income Portion %
0	0
20	20
40	40
80	80
100	100

Let us now see how a graph for this data actually looks:



4.3.4 LIMITATIONS OF GRAPHIC PRESENTATION

As we discussed that graphical presentation is an eye catching way to represent data but one should be very careful for constructing it. There are various limitations of graphical presentation of data i.e.

1. Sometimes graphs only provides half information to a common man and deep understanding of graphs require experts.
2. Graphs only provides limited information.
3. Graphical presentation of data is very costly because it involves images, colours and paints.
4. Graphical presentation requires a lot of time and energy.
5. There is high possibilities of error and mistake in the graphical presentation of data.
6. Graphs presents full information of data which cause lack of secrecy.
7. It is very difficult to select a suitable method for graphical presentation of data.

4.4 CHECK YOUR PROGRESS

1. Why Diagrammatic Presentation is better than Tabulation of Data?

2. Which Bar Diagram is used to show two or more characteristics of the Data?

3. Mention the total of the degrees of all the angles formed at the centre of a Circle.

4. A histogram is a graphical representation of _____.

5. A sector diagram is also known as _____.

4.5 SUMMARY

Now, you are able to learn how collected data can be presented through various forms of presentations or visualization i.e. Diagrammatic and Graphical presentations. Moreover, at this point you are able to understand the different forms of diagrams and graphs. You can easily construct diagrams and graphs, and represent data according to your use. You have an idea about the rules of constructions for presentation of data and you are also aware about the merits and demerits of using the presentation method. Thus, you can make presentation of data meaningful,

Comprehensive and purposeful. Further, presentation of data is very useful for every section of society i.e. whether you are a student, a businessmen or a common make. Thus, through understanding presentation of data you will be able to present data in an efficient manner, you can use appropriate method of presentation and easily spread information to others.

4.6 KEYWORDS

Presentation of Data refers to the organization of data into tables, graphs or charts, so that logical and statistical conclusion can be derived from the collected measurement.

A bar chart consists of a set of bars whose heights are proportional to the frequencies that they represent.

Histogram is a set of vertical bars whose areas are proportional to the frequencies of the classes that they represent.

Ogive is a cumulative frequency graphs drawn on natural scale to determine the values of certain factors like median, Quartile, Percentile etc.

Three dimensional diagrams are the diagrams in which three dimension are taken into account. These dimensions are length, width and height. These diagrams are also known as Cubic Diagram and these diagrams may be drawn in the form of cylinders, blocks, spheres, etc.

4.7 SELF-ASSESSMENT TEST

1. Represent the following data in a bar chart.

The amount (in thousands of litres) of petrol sold at a petrol station during a month was

Type of petrol	Leaded	Unleaded	Diesel
Number of litres (x 1000)	45	35	20

2. Use the following raw data of the length (mm) of nails found in packets of 'assorted nails'.

11	48	53	32	28	15	17	45	37	41
55	31	23	36	42	27	19	16	46	39
41	28	43	36	21	51	37	44	33	40
15	38	54	16	46	47	20	18	48	29
31	41	53	18	24	25	20	44	13	45

a) Make a grouped frequency table taking class intervals 10 -14, 15 - 19, etc., and draw a histogram.

b) Make a grouped frequency table taking class intervals 10 - 19, 20 -29, etc., and draw the histogram.

Compare the two representations of the data.

3. Represent the following data by sub-divided percentage rectangular diagram.

Items of Expenditure	Family A		Family B	
	Rs.	Y	Rs.	Y
Food	2000	40	2500	31
Clothing	1000	20	2000	25
House Rent	800	16	1000	13
Fuel and Lighting	400	8	500	6
Miscellaneous	800	16	2000	25

Total	5000	100	8000	100

4. What do you mean by Diagrammatic presentation of data? Explain the major rule for constructing a diagram.
5. What do you understand by Graphical presentation of data? How it is different from Diagrammatic presentation of data?
6. Comparing the two representations, pie chart and pictogram, list some advantages and disadvantages of each.
7. The number of teacher trained for the Senior Secondary School in Botswana are tabulated:

Year	1996	1997	1998	1999
Number trained	81	105	115	184

a) Represent this data in a line graph and comment on the trend.

b) If the trend is continuous what number of Senior Secondary school teachers do you expect to be trained in the year 2000?

8. Display the following data in a pie chart and pictogram.

The type of vehicles coming to a petrol station during one day are tabulated below:

Person cars 26 Lorries 12 Busses 8 Combis 14

9. Get information from your school office about the CBSE result (2019) for the students of Class XII in your school. Draw a bar diagram (showing their aggregate marks classified as 1st division, 2nd division and 3rd division).

10. Here is an exercise for the students of Class XI. Draw a programme to conduct direct personal oral investigation of all the students of your school. Find out which mode of transport they use to come to the school. Present your information in terms of a pie diagram.

4.8 ANSWERS TO CHECK YOUR PROGRESS

1. It makes data more attractive as compared to tabulation and helps in visual comparison.
2. Multiple Bar Diagram

3. 360°

4. A Frequency Distribution

5. Pie Diagram

4.9 REFERENCES/SUGGESTED READINGS

- Bowen, R. (1992) Graph it! Prentice-Hall, New Jersey. [A readable, though fairly basic, guide]
- Cleveland, W. S. (1994) The elements of graphing data. Hobart Press, New Jersey. [An excellent and detailed look at methods of graphical presentation with particular reference to studies of visual perception]
- Tufte, E. R. (1983) The visual display of quantitative information. Graphics Press, Connecticut.
- Francis, A. (2004). *Business mathematics and statistics*. Cengage Learning EMEA.
- Srinivasa, G. (2008). *Business Mathematics and Statistics*. New Age International.
- Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS biology*, *13*(4).
- Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician*, *38*(4), 270-280.
- Altman, D. G., & Bland, J. M. (1996). Statistics notes: Presentation of numerical data. *BMJ*, *312*(7030), 572.

Subject: Business Statistics-1	
Course Code: BCOM 302	**Author :** Dr. Pradeep Gupta
Lesson No. : 5	**Vetter:** Dr. B.S. Bodla
MEASURES OF CENTRAL TENDENCY	

Structure

5.0 Learning Objectives

After going through this lesson, you will be able to:

- Understand the concept of Arithmetic mean
- Understand the concept of Median
- Understand the concept of Mode
- Understand the concept of Harmonic mean
- Understand the concept of Geometric mean

5.1 Introduction

Central tendency or 'average' value is the powerful tool of analysis of data that represents the entire mass of data. The word 'average' is commonly used in day to day conversation. For example, we often

talk of average income, average age of employee, average height, etc. An 'average' thus is a single value which is considered as the most representative or typical value for a given set of data. Such a value is neither the smallest nor the largest value, but is a number whose value is somewhere in the middle of the group. For this reason an average is frequently referred to as a *measure of central tendency of central value.* Measures of central tendency show the tendency of some central value around which the data tends to cluster.

Characteristics of a Good Average:

Since an average is a single value representing a group of values, it is desirable that such a value satisfies the following properties:

(i) It should be easy to understand: Since statistical methods are designed to simplify complexity, it is desirable that an average be such that it can be readily understood, otherwise, its use is bound to be very limited.

(ii) It should be simple to compute: Not only an average should be easy to understand but also it should be simple to compute so that it can be used widely. However, though ease of computation is desirable, it should not be sought at the expense of other advantages, i.e. if in the interest of great accuracy, use of more difficult average is desirable one should prefer that.

(iii) It should be based on all the observations: The average should depend upon each and every observation so that if any of the observation is dropped average itself is altered.

(iv) It should be rigidly defined: An average should be properly defined so that it has one and only one interpretation. It should preferably by defined by an algebraic formula so that if different people compute the average for the same figures they all get the same answer (barring arithmetical mistakes).

(v) It should be capable of further algebraic treatment: We should prefer to have an average that could be used for further statistical computation. For example, if we are given separately the figures of average income and number of employees of two or more factories, we should be able to compute the combined average.

(vi) It should not be unduly affected by the presence of extreme values: Although each and every observation should influence the value of the average, none of the observation should influence it unduly. If one or two very small or very large observations unduly affect the average

i.e. either increase its value or reduce its value, the average cannot be really typical of the entire set of data. In the words, extremes may distort the average and reduce its usefulness.

The following are important measures of central tendency which are generally used in business:

5.1.1 Arithmetic Mean

The arithmetic mean (usually denoted by the symbol $\bar{x}$) of a set of observations is the value obtained by dividing the sum of all observations in a series by the number of items constituting the series.

Computation of Arithmetic Mean:

1. Un-grouped Data: If $x_1, x_2, \ldots\ldots\ldots\ldots x_n$ are the n given observations, then their arithmetic mean usually denoted by $\bar{x}$ is given by:

$$\bar{\mathfrak{X}} = \frac{X1 + X2 + X3 \ldots\ldots X4}{N}$$

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

The symbol Σ (Greek letter called Sigma) denotes the sum of n items. In normal use only Σ x is written in place of Σx (i=1...n). However, when the sum is combined to a given range of numbers out of the total, then it becomes necessary to specify. j

Problem 1:

The following gives the marks obtained by 10 students at an examination:

Roll Nos.: 1 2 3 4 5 6 7 8 9 10

Marks

Obtained : 43 48 55 57 21 60 37 48 78 59 Calculate the arithmetic mean.

Solution: Computation of Arithmetic Mean

Roll No.	***Marks obtained (x)***

1	43
2	48
3	55
4	57
5	21
6	60
7	37
8	48
9	78
10	59
Total	Σx=506

Arithmetic Mean= $(\sum \boldsymbol{x})/\boldsymbol{n}$

= 506/10

= 50.6

$\bar{x}$ = 50.6 Ans.

Frequency Distribution: In case of a frequency distribution. The arithmetic mean is given by the following formula:

$$= \frac{f_1x_1+f_2x_2+.....f_nx_n}{f_1+f_1+.....f_n} = \frac{\Sigma f_ix_i}{\Sigma f_i} = \frac{\Sigma f_ix_i}{N}$$

Where N=Σf is the total frequency. The mean value obtained in this manner is sometimes referred as *weighted arithmetic mean,* as distinct from *simple arithmetic mean*.

In case of continuous or grouped frequency distribution, the value of x is taken as the mid-value of the corresponding class.

Problem 2:

From the following data of marks obtained by 50 students of a class, calculated the arithmetic mean:

Marks	*No. of Students*	*Marks*	*No. of Students*
20	8	50	5
30	12	60	6
40	15	70	4

Let the marks be denoted by X and the number of students by f.

Solution:

Marks(x)	*No. of Students (f)*	*fx*
20	8	160
30	12	360
40	15	600
50	5	250
60	6	360
70	4	280
		2010

$$\bar{X} = \frac{\Sigma fx}{N} = \frac{2010}{50} = 40.2 \text{ marks}$$

Hence the average mark is 40.2.

Problem 3: Calculate the mean for the following frequency distribution:

Sales (in Rs. lakh):	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of firms :	6	5	8	15	7	6	3

Solution: Computation of Arithmetic Mean

Sales	*Mid-Value*	*No. of Firms*	*fX*
(in Rs. lakh)	*(X)*	*(f)*	
0-10	5	6	30
10-20	15	5	75

20-30	25	8	200
30-40	35	15	525
40-50	45	7	315
50-60	55	6	330
60-70	65	3	195
		Σf=50	ΣfX=167

$$\text{A.M.} = \frac{\Sigma fx}{N} = \frac{1670}{50} = \text{Rs. } 33.4 \text{ lakhs.}$$

Short-cut Method: When the short-cut method is used arithmetic mean is computed by applying the formula given below:

$$X = A + \frac{\Sigma fd}{N}$$

Where, A = assumed mean and d=deviations from assumed mean (m-A).

Problem 3 will be solved as follows when short-cut method is used :

Mid value (m)	:	5	15	25	35	45	55	65
Deviations (m-A)	:	-30	-20	-10	0	10	20	30
A = 35f	:	6	5	8	15	7	6	3
fd	:	-180	-100	-80	0	70	120	90

Here: A = 35, N = 50, Σ fd = -80,

$$\bar{x} = 35 + \frac{-80}{50}$$

$$= 33.4 \text{ lakhs.}$$

Step-deviation Method: In the step deviation method the only additional point is that in order to simplify calculations we take a common factor from the data and multiply the result by the

common factor. The formula is:

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

Where A = assumed mean; F = frequency; $D' = \frac{(m - A)}{(C)}$;

C= common factor, N = Total number of observations.

The step deviation method is most commonly used formula as it facilitate calculations.

Problem 4:

The following table gives the individual output of 180 female workers at a particular plant during a work. Find out the average output per worker.

Output (in units)	500-509	510-519	520-529	530-539
No. of workers	8	18	23	37
Output (in units)	540-549	550-559	560-569	570-579
No. of workers	47	26	16	6
Solution :				
Mid-value (m)	**Frequency (*f*)**		**D_1=m-534.5 10**	***f*d'**
504.5	8		- 3	-24
514.5	18		- 2	-36
524.5	23		- 1	-23
534.5	37		0	0
544.5	47		1	47
554.5	26		2	52
564.5	16		3	48
574.5	5		4	20
	180			Σfd = 84

$$\text{Average output : } \quad \bar{X} = A + \frac{C}{N} \times \Sigma fd$$

$$= 534.5 + \frac{10}{180} 84$$

$$= 534.5 + 4.67$$

$$= 539.17 \text{ units}$$

Mean of the Combined Series

If a group of n_1 observations has A.M. $\bar{X}_1$ and another group of n_2 observations has A.M. $\bar{X}_2$, then the A.M. ($\bar{X}_{12}$) of the composite group of $n_1 + n_2$ (=n, say) observations is given by

$$\bar{X}_{12} = \frac{n_1 \bar{X}_1 + n_2 X_2.}{n_1 + n_2}$$

where, X_{12} = combined mean of the two series or two groups of data. This can be generalised to any number of groups.

Problem 5 : The mean height of 25 male workers in a factory is 61 cms., and the mean height of 25 female workers in the same factory is 58 cms. Find the combined mean height of 50 workers in the

Solution :

$$X_{12} = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$$

$$n_1 = 25, \bar{x}_1 = 61, n_2 = 25\ \bar{x}_2 = 58$$

$$\bar{x}_{12} = \frac{25x61 + 25x58}{50+50} = \frac{1525 + 1450}{50} = \frac{2975}{50} = 59.6$$

Thus combined mean height of 50 workers is 59.5 cms.

Merits and Limitations of Arithmetic Mean

The arithmetic mean is the most popular average in practice. It is due to the fact that it possesses first five out of six characteristics of a goods average (as discussed earlier) and no other average possesses such a large number of characteristics.

However, arithmetic mean is unduly affected by the presence of extreme values. Also in open-end frequency distribution it is difficult to compute mean without making assumption regarding the size of the class-interval of the open-end classes.

Mathematical Properties of Arithmetic Mean

The following area a few important mathematical properties of the arithmetic mean.

1. The sum of the deviations of the items form the arithmetic mean (taking signs into account) is always zero. i.e. $\Sigma(x-x) = 0$.
2. The sum of the squared deviations of the items from arithmetic mean is minimum, that is, less than the sum of the squared deviations of the items from any other value.
3. If we have the arithmetic mean and number of items of two or more than two related groups, we can compute combined average of these groups.

5.1.2 MEDIAN

In the words of L.R. Conner : "The median is that value of the variable which divides the data in two equal parts, one part comprising all the values greater and the other, all values less than median." Thus, as against arithmetic mean which is based on all the items of the distribution, the median is only positional average, i.e. the value depends on the position occupied by a value in the frequency distribution.

Computation of Median

1. **Ungrouped data :** If the number of observation is odd, then the median is the middle value after the observations have been arranged in ascending or descending order of magnitude. In case of even number of observations median is obtained as the arithmetic mean of two middle observations after they are arranged in ascending or descending order of magnitude.

Problem 6: The marks obtained by 12 students out of 50 are: 25, 20, 23, 32, 40, 27, 30, 25, 20, 10, 15, 41

Solution: The values obtained by 12 students arranged in ascending order as: 10, 15, 20, 20, 23,

25, 25, 27, 30, 32, 40, and 41

Here the number of items 'N' = 12, which is even

∴ The two middle items are 6th and 7th items

$$\text{i.e.} \quad \frac{25+25}{2} = 25 \text{ is the median value.}$$

2. Frequency (Discrete) Distribution:

In case of frequency distribution where the variables take the value $X_1, X_2, \ldots\ldots,$ ΣX with respective frequencies $f_1, f_2, \ldots\ldots, f_n$ with $N = f$, median is the size of the $\frac{1}{2}(N+1)$ th item or observation. In this case the use of cumulative frequency (c.f.) distribution facilitates the calculations. The steps involved are:

(i) Prepare the less than cumulative frequency (c.f.) distribution.

(ii) Find N/2.

(iii) Find the c.f. just greater than N/2.

(iv) The corresponding value gives the median.

Problem 7: From the following data find the value of median:

Income (Rs.)	1000	1500	800	2000	2100	1700
No. of Persons	24	26	14	10	5	28
Solution:						
Income arranged in ascending order		***No. of persons (f)***				***c.f.***
800		14				14
1000		24				38
1500		26				64
1700		28				92
2000		10				102
2100		5				107

Median = Size of (N/2)th item = $\frac{107}{2}$ = 53.5

53.5th item is consisted in the c.f. = 64. The corresponding value to this = 1500. Hence Median = Rs. 1500.

3. Continuous Frequency Distribution : Steps involved for its computation are :

(i) Prepare less than cumulative frequency (c.f.) distribution.

(ii) Find N/2.

(iii) Locate c.f. just greater than N/2.

(iv) The corresponding class contains the median value and is called the median class.

(v) The value of median is now obtained by using the interpolation formula :

$$\text{Median (Md)} = 1 + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

Where 1 is the lower limit or boundary of the median class; f is the frequency of the median class; h is the magnitude or width of class interval; n = Σf is the total frequency; and C is the cumulative frequency of the class preceding the median class.

Problem 8: The annual profits (in Rs. lacs) shown by 60 firms are given below:

Profits:	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65
No. of firms:	4	5	11	6	5	8	9	6	4	2

Calculate the median.

Solution :

Profits	***No.of firms (f)***	***Cumulative frequency (c.f.)***
15-20	4	4
20-25	5	9
25-30	11	20
30-35	6	26
35-40	5	31
40-45	8	39

45-50	9	48
50-55	6	54
55-60	4	58
60-65	2	60

Median item = $\frac{1}{2}$ N = 30

The cumulative frequency just greater than 30 is 31 and is corresponding class 35-40 is the median class.

$$\text{Median} = L + \frac{N/2 - c.f.}{f} \times h$$

$$= 35 + \frac{30-26}{5} \times 5 = 39 \text{ marks.}$$

Merits and Limitations of Median

The median is superior to arithmetic mean in certain aspects. For example, it is especially useful in case of open-ended distribution and also it is not influenced by the presence of extreme values. In fact when extreme values are present in a series, the median is more satisfactory measure of central tendency than the mean.

However, since median is positional average, its value is not determined by each and every observation. Also median is not capable of algebraic treatment. For example, median cannot be used for determining the combined median of two or more groups. Furthermore, the median tends to be rather unstable value if the number of observations is small.

5.1.3 MODE

Mode is the value which occurs most frequently in the set of observations.

Computation of Mode

(a) Ungrouped Data: In case of a frequency distribution, mode is the value of the variable corresponding to the maximum frequency.

Problem 9: Calculate the mode of the following data:

Sr. No.	*Marks obtained*	*Sr. No.*	*Marks obtained*
1	16	6	27
2	27	7	20
3	24	8	18
4	12	9	15
5	27	10	15

Solution. : Calculation of Mode

Size of item (Marks)	*No. of times it occurs*	*Size of item (Marks)*	*No. of times it occurs*
12	1	20	1
15	2	24	1
16	1	27	3
18	1		

Since the item 27 occurs the maximum number of times i.e. 3, hence the modal marks are 27.

(b) Grouped Data: From the grouped frequency distribution, it is relatively difficult to find the mode accurately. However, if all classes are of equal width, mode is usually calculated by the formula :

$$\text{Mode} = L + \frac{A_1}{A_1 + A_2} \times h$$

Where, L = the lower limit or boundary of the modal class;

h = magnitude or width of the modal class'

$A_1 = f_1 - f_0$, $A_2 = f_1 - f_2$;

f_1 = frequency of the modal class;

f_0 = frequency of the class preceding the modal class; and f_2 = frequency of the class succeeding the modal class.

Mode is generally abbreviated by the symbol M_0.

The above formula takes the following form:

$$\text{Mode (Mo)} = L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times h = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h \quad (1)$$

Problem 10 : Calculate mode from the following data :

Marks	***No. of students***	***Marks***	***No. of students***
Above 0	80	Above 60	28
Above 10	77	Above 70	16
Above 30	65	Above 80	10
Above 40	55	Above 90	8
Above 50	43	Above 100	0

Solution:

Since this is cumulative frequency distribution, we are to first convert it into a simple frequency distribution.

Marks	***No. of students***	***Marks***	***No. of students***
0-10	3	50-60	15
10-20	5	60-70	12
20-30	7	70-80	6
30-40	10	80-90	2
40-50	12	90-100	8

By inspection the modal class is 50-60.

$$\text{Mode} = L + \frac{A_1}{A_1 + A_2} \times i$$

$L = 50$, $A_1 = (15-12) = 3$, $A_2 = (15-12) = 3$, $i = 10$

$M_0 = 50 + \frac{\text{-----------------------}}{3+3} \text{x } 10 = 50+5 = 55$ Marks.

Empirical Relation between Mean (X), Median (Md) and Mode (M_0)

In case of a symmetrical distribution mean, median and mode coincide i.e. Mean = Median = Mode. However, for a moderately asymmetrical (non-symmetrical) distribution, mean and mode usually lie on the two ends and median lies in between they and they obey the following important empirical relationship, given by Prof. Karl Pearson.

Mode = 3 Median - 2 Mean ----------(2).

While applying the formula (1) for calculating mode, it is necessary to see that class intervals are uniform throughout. If they are unequal they should first be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise we will get misleading results.

A distribution having only one mode is called unimodel. If it contains more than one mode, it is called bimodal or multimodal. In the latter case the values of the mode cannot be determined by formula (1) and hence mode is ill-defined when there is more than one value of mode. Where mode is ill-defined, its value is ascertained by the formula (2) based upon the relationship between mean, median and mode. Mode = 3 Median - 2 Mean.

Merits and Limitations of Mode

Like Mean, the mode is not affected by extreme values and its value can be obtained in open-end distribution without ascertaining the class limits. Mode can be easily used to describe qualitative phenomenon. For example, when we want to compare the consumer preferences for different types of products, say, soap, toothpastes etc. or different media of advertising, we should compare the modal preferences. In such distributions where there is an outstanding large frequency, mode happens to be meaningful as an average.

However, mode is not rigidly defined measure as there are several formulae for calculating the mode, all of which usually give somewhat different answer. Also the value of mode cannot always be computed, such as, in case of binomial distributions.

5.2 GEOMETRIC MEAN

The Geometric mean (usually abbreviated as G.M.) of a set of n observations is the nth root of their product.

Computation of Geometric Mean

The Geometric Mean G.M. of n observations X_i, i=1, 2,……, n is G.M. = $(X_1.X_2.....X_n)^{1/n}$

The computation is facilitated by the use of logarithms. Taking logarithms of both sides, we get.

$$\log G.M. = \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n) = \frac{1}{n}\sum_{i=1}^{n}\log x_i$$

$$G.M. = \text{Antilog}\left(\frac{1}{n}\sum_{i=1}^{n}\log x_i\right) \text{ or } \text{antilog}\left(\frac{1}{n}\sum \log\right)$$

Problem 11: From the data given below calculate the G.M.

15, 250, 15.7, 157, 1.57, 105.7, 10.5, 1.06, 25.7, 0.257

Solution:

Value (x)	*Log (x)*
15	1.1761
250	2.3979
15.7	1.1959
157	2.1959
1.57	0.1959
105.7	2.0240
10.5	1.0212
1.06	0.0253
25.7	1.4099
0.257	0.0409
Total	**11.0520**

G.M. = Antilog $\left(\frac{1}{n}\Sigma \log x\right)$

$$\text{G.M.} = \text{Antilog}\left(\frac{11.0520}{10}\right) = 12.75$$

In case of frequency distribution x_i/f_i (i=1, 2,..... , n)geometric mean, G.M. is given by

$$\text{G.M.} = n\sqrt{(x_1.x_1 \ldots\ldots f_1 \text{ times})\,(x_2.x_2 \ldots\ldots f_2 \text{ times}) \ldots\ldots (x_n.x_n \ldots\ldots f_n \text{ times})}$$

Since the product of the values in a frequency distribution is usually very large, formula (3) is not suitable in computing the value of G.M. Taking logarithm of both sides in (3), we get :

$$\log \text{G.M.} = \frac{1}{N}\{\log (x_1^{f1}.x_2^{f2} \ldots\ldots x_n^{fn})\}$$

$$= \frac{1}{N}\{f_1 \log x_1 + f_2 \log x_2 + \ldots\ldots\ldots\ldots + f_n \log x_n)$$

Problem 12: Calculate Geometric Mean of the following distribution.

X	:	70	100	103	107 149	
f	:	10	12	8	5 5	
Solution :						
	X		***f***		***log x***	***f log x***
	70		10		1.8451	18.4512
	100		12		2.0000	24.0000
	103		8		2.0128	16.1024
	107		5		2.0294	10.1470
	149		5		2.1732	10.8690
						79.5664

$$\text{Log G.M.} = \frac{\Sigma f \log x}{\Sigma f}$$

$$= \frac{79.5664}{40}$$

$$= 1.989$$

In the case of grouped frequency distribution, the values of x are the mid-values of the corresponding classes.

Combined Geometric Mean

Just as we have talked of combined arithmetic mean, in a similar manner we can also talk of combined geometric mean. If the Geometric mean of N observations is G.M. and these observations are divided into two sets containing N_1 and second containing N_2 observations having GM_1 and GM_2 as the respective geometric means, then:

$$\log GM = \frac{N_1 \log GM_1 + N_2 \log GM_2}{N_1 + N_2}$$

Merits and Limitations of Geometric Mean

Geometric mean is highly useful in averages, ratios, percentages and in determining rates of increase and decrease. It is also capable of algebraic manipulation. For example, if the geometric mean of two or more series and their number of observations are known, a combine geometric mean can easily be calculated.

However, compared to arithmetic mean, this average is more difficult to compute and interpret. Also geometric mean cannot be computed when odd number of observations is negative.

5.3 HARMONIC MEAN

Harmonic mean of a number of observations is the reciprocal of arithmetic mean of reciprocals of the given values.

Computation of Harmonic Mean: If $X_1, X_2, \ldots\ldots\ldots\ldots, X_n$ are the n observations, their harmonic mean (abbreviated as H) is given by :

$$\text{H.M. (H)} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots\ldots + \frac{1}{x_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Problem 13: Find the Harmonic mean from the following:

2574, 475, 75, 5, 0.8, 0.08, 0.005, 0.0009

Solution:

X	***1/x***	***X***	***1/X***
2574	0.0004	0.8	1.2500
475	0.0021	0.08	12.5000
75	0.0133	0.005	200.0000
5	0.2000	0.0009	1111.1111

$\Sigma(1/x)=1325.0769$

$$\text{H.M.} = \frac{n}{(1/x)\ \Sigma} = \frac{8}{1325.0769} = 0.006$$

In case of frequency distribution, we have

$$\frac{1}{H} = \frac{1}{N}\left[\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots\ldots\ldots + \frac{f_n}{x_n}\right] \text{ where } N=\Sigma f$$

Problem 14: The following table gives weights of 31 persons in a sample enquiry. Calculate mean by using Harmonic mean.

Weight (in 1bs): 130	135	140	145	146	148	149	150	157
No. of persons: 3	4	6	6	3	5	2	1	1

Solution:

(Weight(x)	***Frequency(f)***	***1/x***	***f(1/x)***
130	3	.00769	.02307
135	4	.00741	.02964
140	6	.00714	.04284
145	6	.00690	.04140
146	3	.00685	.02055

148	5	.00676	.03380
149	2	.00671	.01342
150	1	.00667	.00667
157	1	.00637	.00637
	31		.21776

$$\frac{1}{\text{H.M.}} = \frac{\text{Sf. } 1/x}{N} = \frac{.21776}{31} = .007024$$

$$\text{Or H.M.} = \frac{1}{.007024} = 142.4 \text{ lbs.}$$

The harmonic mean is restricted in its field of application. The harmonic mean is a measure of central tendency for data expressed as rates, for instance - kms. per hour, tonnes per day, kms per litre etc.

Merits and Limitations of Harmonic Mean

The harmonic mean, like the arithmetic mean and geometric mean is computed from all observations. It is useful in special cases for averaging rates. However, harmonic mean gives largest weight to smallest observations and as such is not a good representation of a statistical series. In dealing with business problems harmonic mean is rarely used.

5.4 Check Your Progress

There are some activities to check your progress. Answer the followings:

1. The sum of the squared deviations of the items from arithmetic mean is…….
2. Harmonic mean gives largest weight to ……observations and as such is not a good representation of a statistical series.
3. Geometric mean cannot be computed when odd number of observations is…….
4. Mode can be easily used to describe ……..phenomenon.
5. In case of a ……..distribution mean, median and mode coincide.

5.5 Summary

It is the most important objective of statistical analysis is to get one single value that describes the characteristics of the entire mass of cumbersome data. Such a value is finding out, which is known as

central value to serve our purpose. Central tendency or 'average' value is the powerful tool of analysis of data that represents the entire mass of data. An 'average' thus is a single value which is considered as the most representative or typical value for a given set of data. Such a value is neither the smallest nor the largest value, but is a number whose value is somewhere in the middle of the group. Measures of central tendency show the tendency of some central value around which the data tends to cluster. There are different measure for central tendency like mean, median, mode, harmonic mean and geometric mean. There are various merits and limitations of each and having different characteristics.

5.6 Keywords

Average: It is a single value which is considered as the most representative or typical value for a given set of data.

Mean: It a set of observations is the value obtained by dividing the sum of all observations in a series by the number of items constituting the series.

Median: It is that value of the variable which divides the data in two equal parts, one part comprising all the values greater and the other, all values less than median.

Mode: It is the value which occurs most frequently in the set of observations.

5.7 Self- Assessment Test

Q1. What is the measures of central tendency? Why are they called measures of central tendency?

Q2. What is the properties of a good average?

Q3. Give a brief note of the measures of central tendency together with their merits and demerits. Which is the best measured of central tendency and why?

Q4. Following distribution gives the pattern of overtime work done by 100 employees of a company. Calculate median.

Overtime Hours:	10-15	15-20	20-25	25-30	30-35	35-40
No. of Employees:	11	20	35	20	8	6

Q5. The mean monthly salary paid to all employees in a company is Rs. 1600. The mean monthly salaries paid to technical and non-technical employees are Rs. 1800 and Rs. 1200 respectively. Determine the percentage of technical and non-technical employees of the company.

Q6. Calculate the arithmetic mean and the median of the frequency distribution given below. Also calculate

the mode using the empirical relation among the three:

Class Limits	*Frequency*	*Class Limits*	*Frequency*
130-134	5	150-154	17
135-139	15	155-159	10
140-144	28	160-164	1
145-149	24		

Q7. In a certain factory a unit of work is completed by A in 4 minutes, by B in 5 minutes, By C in 6 minutes, by D in 10 minutes and by E in 12 minutes.

(a) What is the average number of units of work completed per minute?

(b) At this rate how many units will they complete in a six-hour day?

Q8. Find the average rate of increase in population which in the first decade increased by 20%, in the second decade by 30% and in the third decade by 40%.

Q9. In a class of 50 students, 10 has failed and their average of marks is 2.5. The total marks secured by the entire class were 281. Find the average marks of those whohave passed.

5.8 Answers to Check your Progress

1. Minimum
2. Smallest
3. Negative
4. Qualitative
5. Symmetrical

5.9 References/Suggested Readings:

- Gupta, S. P.: Statistical Methods, Sultan Chand and Sons, New Delhi.
- Levin, R. I. and David, S. R.: Statistics for Management, Prentice Hall, New Delhi.
- Gupta, C. B.: Introduction to Statistical Methods.
- Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.

Subject: Business Statistics-1	
Course Code: BCOM 302	**Author :** Dr. Pradeep Gupta
Lesson No. : 6	**Vetter:** Dr. B.S. Bodla
MEASURES OF DISPERSION	

Structure

6.0 LEARNING OBJECTIVES

After going through this lesson, you will be able to:

- Understand the concept of dispersion
- Understand the concept of Range

- Understand the concept of Quartile Deviation
- Understand the concept of Mean Deviation
- Understand the concept of Standard Deviation
- Understand the concept of coefficient of Variance
- Explain the concept of Lorenz Curve

6.1 INTRODUCTION

The value given by a measure of central tendency is considered to be the representative of the whole data. However, it can describe only one of the important characteristics of a series. It does not give the spread or range over which the data are scattered. Measures of dispersion are used to indicate this spread and the manner in which data are scattered.

6.1.1 DEFINITION

Dispersion indicates the measure of the extent to which individual items differ from some central value. It indicates lack of uniformity in the size of items. Some important definitions of dispersion are given below:

(1) According to Spiegel, "The degree to which numerical data tend to spread about an average value is called the variation of dispersion of the data."

(2) Simpson and Kafka define dispersion as "The measurement of the scatterness of the mass of figures in a series about an average is called measure of variation or dispersion."

(3) As defined by Brooks and Dick, "Dispersion or spread is the degree of the scatter or variation of the variable about a central value."

Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of the *second order*.

6.1.2 USES OF MEASURES OF DISPERSION

Average is a typical value but it alone does not describe the data fully. It does not tell us how the items in a series are scattered around it. To clear this point considers the following three sets of data:

Set A	30	30	30	30	30
Set B	28	29	30	31	32
Set C	3	5	30	37	75

All the three sets A, B and C have mean 30 and median is also 30. But by inspection it is apparent that the three sets differ remarkably from one another. Thus to have a clear picture of data, one needs to have a measure of dispersion or variability (scatteredness) amongst observations in the set. It is also used for comparing the variability or consistency (uniformity) of two or more series. A higher degree of variation means smaller consistency.

6.1.3 PROPERTIES OF A GOOD MEASURE OF DISPERSION

There are various measures of dispersion. The difficulty lies in choosing the best measure as no hard and fast rules have been made to select any one. However, some norms have been set which work as guidelines for choosing a particular measure of dispersion. A measure of dispersion is good or satisfactory if it possesses the following characteristics.

- It is easily understandable.
- It utilizes all the data.
- It can be calculated with reasonable ease and rapidity.
- It affords a good standard of comparison.
- It is suitable for algebraic and arithmetical manipulation.
- It is not affected by sampling variations.
- It is not affected by the extreme values.

6.1.4 VARIOUS MEASURES OF DISPERSION

Commonly used measures of dispersion are:

(a) Range

(b) Quartile deviation

(c) Mean deviation

(d) Standard deviation

(a) Range

Definition. Range is the difference between the two extreme items, i.e. it is the difference between the maximum value and minimum value in a series.

Range (R) = Largest value (L) minus Smallest value (S) A relative measure known as *coefficient of range is* given as:

$$\text{Coefficient of range} = \frac{L - S}{\quad}$$

$$L + S$$

Lesser the range or coefficient of range, lower the variability.

Properties.

(a) It is the simplest measure and can easily be understood.

(b) Besides the above merit, it hardly satisfies any property of a good measure of dispersion e.g. it is based on two extreme values only, ignoring the others. It is not liable to further algebraic treatment.

Example 1. The population (in '000) in eighteen Panchayat Samities of a district is as given below:

77, 76, 83,	68,	57,	107, 80,	75,	95,	100, 113, 119,	
121, 121, 83,	87,	46,	74				

Calculate the range and coefficient of range.

Solution. Largest value (L)		=	121
Smallest value (S)		=	46
Range (R)		=	L - S
		=	121-46 = 75
Coefficient of range = $\frac{L - S}{L + S}$	=	$\frac{121-46}{121+46}$ = $\frac{75}{167}$ = 0.449	

Range for grouped data. In case of grouped data, the range is the difference between the upper limit of the highest class and the lower limit of the lowest class. No consideration is given to frequencies.

Example 2. Find range of the following distribution.

Class-interval	45-49	50-54	55-59	60-64	65-69
Frequency	37	26	8	5	1

Solution. The series can be written as follows :

Group	Frequency
44.5-49.5	37
49.5-54.5	26

54.5-59.5	8
59.5-64.5	5
64.5-69.5	1
Range = 69.5-44.5 = 25	

(b) Quartile Deviation

Quartile deviation is obtained by dividing the difference between the upper quartile and the lower quartile by 2.

$$\text{Quartile deviation or Q.D.} = \frac{\text{Upper Quartile - Lower Quartile}}{2}$$

$$= \frac{Q_3 - Q_1}{2}$$

The coefficient of quartile deviation is given by the following formula:

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 - Q_1}$$

Coefficient of quartile deviation is a relative measure of dispersion and is used to compare the variability among the middle 50 per cent observations.

Properties.

(i) It is better measure of dispersion than range in the sense that it is based on the middle 50 per cent observations of a series of data rather than only two extreme values of a series.

(ii) It excludes the lowest and the highest 25% values.

(iii) It is not affected by the extreme values.

(iv) It can be calculated for the grouped data with open end intervals.

(v) It is not capable of further algebraic treatment.

(vi) It is not considered a good measure of dispersion as it does not show the scattering of the central value. In fact it is a measure of partitioning of distribution. Hence it is not commonly used.

Example 3. Given the number of families in a locality according to monthly per capita expenditure classes in rupees as:

Class-interval	140-150	150-160	160-170	170-180	180-190	190-200
No. of families	17	29	42	72	84	107

200-210	2 10-220	220-230	230-240	240-250
49	34	31	16	12

Calculate Quartile deviation and coefficient of quartile deviation.

Solution.

Monthly per capita expenditure (Rs.)	Number of Families (f)	Cumulative frequency (c.f.)
140-150	17	17
150-160	29	46
160-170	42	88
170-180	72	160
180-190	84	244
190-200	107	351
200-210	49	400
210-220	34	434
220-230	31	465
230-240	16	481
240-250	12	493

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

(i) To calculate Q_1, we have to first find :

$$\frac{N}{4} = \frac{493}{4} = 123.25$$

The number 123.25 is contained in the cumulative frequency 160. Hence the first quartile lies in the class 170-180. By using the formula for Q_1 we get,

$$Q_1 = L + \frac{N/4 - c.f.}{f} * i$$

L = 170, N/4 = 123.25, c.f. = 88, f=72, i=10

$$Q_1 = 170 + \frac{123.25 - 88}{72*10}$$

= Rs. 174.90

(ii) To calculate Q_3 we find:

$$\frac{3N}{4} \qquad \frac{3 \times 493}{4}$$

= 369.75

The number 369.75 is contained in the cumulative frequency 400. Hence the class 200-210 is the third quartile class. By using the formula for Q_3 we get:

Q_3=L+ (3 N/4 - c.f.)/f * i
L= 200, 3 N/4 = 369.75, c.f. = 351, f=49, i=10
Q_3 = 200+ (369.75-351)/49 * 10

= Rs. 203.83

Quartile Deviation (Q.D.) = (203.83-174.90)/2=28.93/2=14.465

Coefficient of Q.D = (203.83-174.90)/ (203.83+174.90)=28.93/378.73=0.076

(C) Mean Deviation

Mean deviation is the mean of deviations of the items from an average (mean, median or mode). As we are concerned with the deviations of the different values from an average and in finding the mean of deviations, we have to find the sum of deviations (whether positive or negative); we take all the deviations as positive. We are concerned with the deviations and not with their algebraic signs. We ignore negative signs because the algebraic sum of the deviations of individual values from the average is zero.

Calculation of mean deviation (M.D.).

Mean deviation of a set of n observations $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ is calculated as follows:

$$M.D. = \frac{1}{n} \sum_{i=1}^{n} |x_i - A|$$

for i= 1,2, ..., n where A is a central value.

Let $|x_i - A| = d_i$

Then $M.D. = \frac{1}{n} \sum_{i=1}^{n} |d_i|$ ..... (i)

In case data is given in the form of a frequency distribution, the variate values $x_1, x_2, \ldots\ldots, x_n$ occur $f_1, f_2, \ldots, f_n$ times respectively.

In such series the formula for mean deviation is,

$$M.D. = \frac{1}{N} \sum_{i=1}^{n} f_i |x_i - A|$$ ..... (ii)

Where, $N = \Sigma f_i$ for i = 1, 2, ..., n

In case of grouped data, the mid-point of each class interval is treated as x_i and we can use the formula (ii) given above.

Properties.

(i) Mean deviation removes one main objection of the earlier measures, that it involves each value of the set.

(ii) Its main drawback is that algebraic negative signs of the deviations are ignored which is mathematically unsound.

(iii) Mean deviation is minimum when the deviations are taken from median.

(iv) It is not suitable for algebraic treatment.

Example. 4 :

Calculate mean deviation from the mean for the following data:

Size (x) :	2	4	6	8	10	12	14	16
Frequency :	2	2	4	5	3	2	1	1

Solution :

X	f	F_x	$\|x-8\|$ $\|D\|$	$f\|d\|$
2	2	4	6	12
4	2	8	4	8
6	4	24	2	8
8	5	40	0	0
10	3	30	2	6
12	2	24	4	8
14	1	14	6	6
16	1	16	8	8
	N=20	Σ fx=1600		Σf D = 56

$$X = \frac{\Sigma fx}{N} = \frac{160}{20} = 8$$

$$M.D. = \frac{\Sigma f|D|}{N} = \frac{56}{20} = 2.8$$

Examples 5. Calculate the mean deviation (using median) from the following data.

Size of items	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Solution:

Size	Frequency (f)	Cummulative frequency	Deviation from median 9 |d|	f|d|
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12
			$\Sigma f\|d\| =$	60

$$\text{Median} = \text{Size of } \frac{48+1}{2} \text{ th item}$$

$$= \text{Size of } 24.5^{th} \text{ item} = 9$$

$$\text{Mean deviation} = \frac{\Sigma f|d|}{N} = \frac{60}{48} = 1.25$$

Example. 6 : Find the median and mean deviation of the following data :

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

Solution : Calculation of Median and Mean Deviation.

Size	***f***	***c.f.***	***m.p.***	***| m-35.2 |***	***f |D|***

				$\lvert D \rvert$	
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			$\Sigma f D$ = 1314.8 ~ 1315	

Median = Size of N/2 th item = 100/2 = 50th item.

Median lies in the class 30-40.

$$\text{Med.} = L + \frac{N/2 - \text{c.f.}}{f} \times i$$

L = 30, N/2 = 50, c.f. = 37, f= 25, i = 10.

$$\text{Med.} = 30 + \frac{50-37}{25} \times 10 = 30 + 5.2 = 35.2$$

$$\text{M.D.} = \frac{\Sigma f D}{N} = \frac{1315}{100} = 13.15$$

Uses of Mean Deviation:

The outstanding advantage of the average deviation is its relative simplicity. It is simple to understand and easy to compute. Anyone familiar with the concept of the average can readily appreciate the meaning of the average deviation. If a situation requires a measure of dispersion that will be presented to the general public or any group not familiar with statistics, the average deviation is useful.

(D) Standard deviation

It is the square root of the quotient obtained by dividing the sum of squares of deviations of items from the Arithmetic mean by the number of observations.

$$\therefore \text{Standard deviation or } \sigma = \sqrt{\frac{\text{Sum of squares of deviation from A.M.}}{\text{Number of observations}}}$$

Standard deviation is an absolute measure of dispersion.

Calculation of standard deviation (σ).

(a) Ungrouped data

First method : $\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma d^2}{n}}$

Where d is the deviation of value from the mean.

Second method : In this method, we assume a provisional mean and find the deviations of the values from the provisional mean. The following formula is applied under this method :

$$\text{S.D. or } \sigma = \sqrt{\frac{\Sigma d_x^{\,2}}{n} - \left[\frac{\Sigma d_x}{n}\right]^2}$$

Where d is the deviation of values of x observations from the assumed mean. This formula is more useful when values are in decimals and the mean of the series does not happen to be an integer.

In case the frequencies are also given, then standared deviation is calculated by using the formula :

$$\text{S.D. or } \sigma = \sqrt{\frac{\Sigma f d_x^{\,2}}{n} - \left[\frac{\Sigma f d_x}{n}\right]^2} \qquad \text{Where } n = \Sigma f$$

Example 7. Compute the standard deviation by the short method for the following data :

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution. Let us assume that mean is 15.

X	d (x-15)	d^2
11	- 4	16
12	-3	9
13	-2	4
14	-1	1
15	0	0
16	1	1
17	2	4
18	3	9
19	4	16
20	5	25
21	6	36
	Σ d=11	Σ d2 = 121

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{121}{11} - \left[\frac{11}{11}\right]^2}$$

$$= \sqrt{11 - 1}$$

$$= \sqrt{10} = 3.16$$

In continuous series, we take the central values of the groups.]

Example 8. : Find the standard deviation of the following distribution :

Age	:	20-25	25-30	30-35	35-40	40-45	45-50
No. of Persons	:	170	110	80	45	40	35

Take assumed average = 32.5

Solution :

Calculation of Standard Deviation.

Age m.p. No. of persons (m-32.5)/5

	M	***f***	***d***	***fd***	***fd^2***
20-25	22.5	170	- 2	-340	680
25-30	27.5	110	- 1	-110	110
30-35	32.5	80	0	0	0
35-40	37.5	45	1	45	45
40-45	42.5	40	2	80	160
45-50	47.5	35	3	105	315
		N=480		Σ fd = - 220	Σfd^2 = 1310

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N}\right]^2} \times i$$

$$= \sqrt{\frac{1310}{480} - \left[\frac{-220}{480}\right]^2} \times 5$$

$$= \sqrt{2.729 - .21} \times 5 \quad = 1.587 \times 5 = 7.936$$

Uses of the Standard deviation: As a measure of dispersion, standard deviation is most important. By comparing the standard deviations of two or more series, we can compare the degree of variability or consistency. It is a keystone in sampling and correlation and is also used in the interpretation of normal and skewed curves. It is used to gauge the representativeness of the mean also.

Merits of Standard Deviation:

- It is suitable for algebraic manipulation.
- It is less erratic.
- Standard deviation is considered to be the best measure of dispersion and is used widely.

Demerits of Standard Deviation:

- Its calculation demand greater time and labour.
- If the unit of measurement of variables of two series is not the same, then their variability cannot be compared by comparing the values of standard deviation.
- It gives more weight to extreme items and less to those which are nearer the mean. It is because of the fact that the squares of the deviations which are big in size would be proportionately greater than the squares of those deviations which are comparatively small. The deviations 2 and 8 are in the ratio of 1: 4 but their squares i.e. 4 and 64, would be in the ratio of 1: 16.

Mathematical Properties of Standard Deviation

Standard deviation has some very important mathematical properties which considerably enhance its utility in statistical work.

1. *Combined Standard Deviation :* Just as it is possible to compute combined mean of two or more than two groups, similarly we can also compute combined standard deviation of two or more groups.
2. *Standard deviation of n natural numbers :* The standard deviation of the first in natural numbers can be obtained by the following formula :

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

3. The sum of the squares of deviations of items in the series from their arithmetic mean is minimum. This is the reason why standard deviation is always computed from the arithmetic mean.
4. The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are located with the help of Teheycheff 's theorem, given by mathematician P.L. Tehebycheff (1821-1894). No matter what the shape of the distribution is, at least 75 percent of the values will fall within $\pm$ 2 standard deviation from the mean of the distribution, and at least 89 percent of values will be within + 3 standard deviations from the mean.

   For a symmetrical distribution, the following relationships hold good:

   Mean $\pm$ 1 o covers 68.27% of the items. Mean $\pm$ 2 o covers 95.45% of the items. Mean $\pm$ 3 o

covers 99.73% of the items.

6.2 Variance

The variance is just the square of the standard deviation value:

Variance = $o^2 = (S.D.)^2$

In a frequency distribution where deviations are taken from assumed mean, variance may directly be computed as follows

$$\text{Variance} = \left\{ \frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N} \right)^2 \right\} \times i$$

Where $d = \frac{x-A}{i}$ and i = common factor.

Properties :

(i) The main demerit of variance is that its value is the square of the unit of measurement of variate values. For example, the variable x is measured in cms, the unit of variance is cm. Generally, this value is large and makes it difficult to decide about the magnitude of variation.

(ii) The variance gives more weightage to the extreme values as compared to those which are near to mean value, because the difference is squared in variance.

(iii) The combined groups without redoing the entire calculations.

(iv) Obviously, the combined standard deviation can be found by taking the square root of the combined variance.

Pooled or combined variance : By the combined variance of two groups, we mean the variance of the observations of the two groups taken together. Let us consider two groups consisting of n_1 and n_2 observations respectively. Suppose the means of the groups are $\overline{x}_1$ and $\overline{x}_2$ and the variances are σ_1^2 and σ_2^2 respectively. We know that the pooled mean of both the groups is,

$$\overline{X}_{12} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}$$

The combined variance of the two groups is given by the formula :

$$\sigma_{12}^2 = \frac{n_1\ (\sigma_1^2 + d_1^2) + n_2\ (\sigma_2^2 + d_2^2)}{n_1 + n_2}$$

Where, $d_1 = (\overline{x}_1 - \overline{x}_{12})$ and $d_2 = (\overline{x}_2 - \overline{x}_{12})$

The advantage of the formula of combined variance is that once we know the individual mean and variance of each group, we can calculate the variance of

Example 9: For a group of 50 male workers, the mean and standard deviation of their weekly wages are Rs. 63 and Rs. 9 respectively. For a group of 40 female workers these are Rs. 54 and Rs. 6, respectively. Find the standard deviation for the combined group of 90 workers.

Solution:

The data is $n_1 = 50$ $\bar{x}_1 = 63$ $\sigma_1 = 9$

$n_1 = 40$ $\bar{x}_2 = 54$ $\sigma_2 = 6$

Combined mean $\bar{x}_{12}$ for group of 90 $= \dfrac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$

$= [50 \times 63 + 40 \times 54] / 90$

$= [3150 + 2160]/90 = 5310/90 = 59$

$$\text{Combined standard deviation} = \frac{n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)}{n_1 + n_2}$$

Where, $d_1 = (\bar{x}_1 - \bar{x}_{12})$ and $d_2 = (\bar{x}_2 - \bar{x}_{12})$

$\sigma^2_{12} = [50 (81+16) + 40 (36+25)] / 90$

$= [97 \times 50 + 40 \times 61] / 90 = [4850 + 2440] /90$

$= 7290 / 90 = 81$

$\therefore \sigma_{12} = 9$

Example 10 : The analysis of the results of a budget survey of 150 families gave an average monthly expenditure of Rs. 120 on food items with a standard deviation of Rs. 15. After the analysis was completed it was noted that the figure recorded for one household was wrongly taken as Rs. 15 instead of Rs.105. Determine the correct value of the average expenditure and its standard deviation.

Solution.

A.M. (x) = Rs. 120, No. of items = 150 Total as obtained = 120 x 150 = 18000

Correct total = (Total obtained - item misread) + correct item

= (18000-15) + 105 = 18,090

Correct mean = Correct total/No. of items = 18090/150

= Rs. 120.6

$$(\text{S.D.})^2 = \frac{\Sigma x2}{n} - \left[\frac{\Sigma x}{n}\right]^2$$

Before Correction $(15)^2 = \frac{\sum x^2}{150} - (120)^2$

Or $225 = \frac{\sum x^2}{150} - 14400$ or $\frac{\sum x2}{150} = 14,625$

$\sum x^2 = 14,625 \times 150$

Correct sum of squares = Sum of suqares before correction, minus square of misread item plus square of correct item

= 14625x150 - 15 x 15 + 105 x 105

= 15 x 15 [9750-1 + 7x7] = 225

= [9798]

Correct $(S.D.)^2$ = $225x9798/150 - (120.6)^2$

	=	1.5 x 9798 - 14544.36
	=	152.64
Correct S.D.	=	12.4

6.2.1 Coefficient of variation (C.V.)

It two series differ in their units of measurement; their variability cannot be compared by any measure given so far. Hence in situations where either the two series have different units of measurements, or their means differ sufficiently in size, the coefficient of variation should be used as a measure of dispersion. It is a unit less measure of dispersion and also takes into account the size of the means of the two series. It is the best measure to compare the variability of two series or set of observations. A series with less coefficient of variation is considered more consistent.

Definition.

Coefficient of variation of a series of variate values is the ratio of the standard deviation to the mean multiplied by 100. If σ is the standard deviation and $\bar{x}$ is the mean of the set of values, the coefficient of variation is,

$$C.V. = \frac{\sigma}{\bar{x}} \times 100$$

This measure was given by Professor Karl Pearson.

Properties:

(i) It is one of the most widely used measures of dispersion because of its virtues.

(ii) Smaller the value of C.V., more consistent is the data and vice-versa.

Hence a series with smaller C.V. than the C.V. of other series is more consistent, i.e. it possesses variability.

Example 11: A time study was conducted in a factory with the help of two samples A and B consisting of 10 workers. The time taken by the workers in each case recorded. From the particulars given below state which of the samples is more variable and which takes less time on an average. Time taken in minutes.

Sample A	130	125	120	135	140	145	130	145	140	150
Sample B	132	146	137	145	130	125	138	140	143	144

Solution : Let us calculate mean and standard deviation first by rearranging the data in ascending order.

For Sample A For Sample B

X	d = x - 140 / 5	d^2	y	d'= (y-140)	d'^2
120	- 4	16	125	-15	225
125	- 3	9	130	-10	100
130	- 2	4	132	- 8	64
130	- 2	4	137	- 3	9
135	- 1	1	138	- 2	4
140	0	0	140	0	0

140	0	0	143	3	9
145	1	1	144	4	16
145	1	1	145	5	25
150	2	4	146	6	36
	Σd=-8	Σd2=40		Σ d' = -20	Σ d'2 = 488

$$0 = 140 + \left(\frac{-8}{10} \times 5\right) = 136 \qquad y = 140 + \left(\frac{-20}{10}\right) = 138$$

This shows that on an average workers from sample A takes less time.

$$\sigma_x^2 = \left\{\frac{\Sigma d2}{n} - \left(\frac{\Sigma d}{n}\right)^2\right\} \times i^2 \text{ where } i = 5$$

$$= 25 \left\{\frac{40}{10} - \left(\frac{8}{10}\right)^2\right\} = 25 \left\{\frac{16}{25}\right\} = 84$$

$$\therefore CV_x = \frac{\sigma x}{\overline{X}} \times 100 = \frac{\sqrt{84}}{136} \times 100 = 6.75\ \%$$

Similarly,

$$\sigma_y^2 = \left\{\frac{\Sigma d'2}{n} \left(\frac{\Sigma d'}{n}\right)^2\right\}$$

$$= \left\{\frac{488}{10} -- \left(\frac{-20}{10}\right)^2\right\}$$

$$= 48.8 - 4 = 44.8$$

$$\therefore CV_y = \frac{\sigma y}{\overline{y}} \times 100 = \frac{44.8}{138} \times 100 = 4.85\ \%$$

Since $CV_x > CV_y$, the sample 'A' is more variable, though as we have seen the workers of sample A takes less time.

6.2.2 RELATION BETWEEN STANDARD DEVIATION, MEAN DEVIATION AND QUARTILE DEVIATION

In any bell-shaped distribution, the S.D. will always be larger than M.D. and M.D. larger than Q.D. if the distribution approximates the form of normal curve, the M.D. will be 4/5 of S.D. and Q.D. will be about 2/3rd as large as S.D. Usually,

M.D. = 4/5 of S.D.

Q.D. = 2/3 of S.D.

6.2.3 COMPARISION OF THE VARIOUS MEASURES OF DISPERSION

Range is not a very satisfactory measure of dispersion because it depends solely on the two extreme values and may be very misleading if there are one or two abnormal items. It is impossible to estimate the range in case where there are open ends series. Therefore, it is an unreliable measure of dispersion. Quartile deviation is most easy to calculate and interpret but it is not amenable to mathematical treatment. Mean deviation is easy to compute but grouped data may be difficult. In almost all other aspects, the advantage rests with the standard deviation. Only the S.D. is suitable for algebraic manipulation. For this reason, it is used in correlation, in sampling and in other aspects of advanced statistics. We can compute the S.D. of the whole group if means and standard deviations of two or more subgroups are known. When it is required to compare two or more than two series or distributions, we compute relative measure of dispersion.

6.3 Lorenz Curve

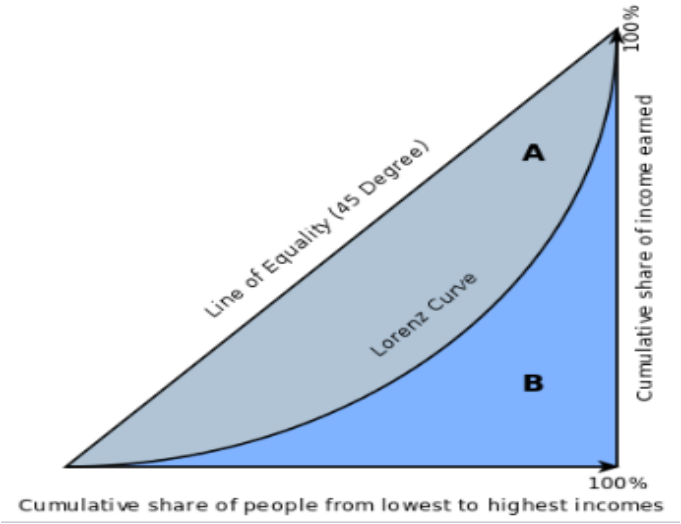
A Lorenz curve is a graph used in economics to show **inequality in income spread or wealth**. It was developed by Max Lorenz in 1905, and is primarily used in economics. However, it may also be used to show inequality in other systems. The Gini index (The **Gini coefficient** is a statistic which quantifies the amount of inequality that exists in a population. The Gini coefficient is a number between 0 and 1, with 0 representing perfect equality and 1 perfect inequality) can be calculated from a Lorenz curve by taking the integral of the curve and subtracting from 0.5.

Reading a Lorenz curve

The x-axis on a Lorenz curve typically shows the portion or percentage of the total population and the y-axis shows the portion of total income/ wealth, or whatever is being analyzed. Since perfect equality would mean that a 1/k portion of the population controlled 1/k of the wealth, perfect equality on this

graph would be shown by a straight line with a slope of 1. This line is often drawn on the graph as a point of reference, alongside the curved line which represents the actual wealth/income/size distribution. The further away from the 1/1 baseline a particular curve is, the more pronounced the inequality. Any point on the curve can be read to tell us what percentage or portion of the population command what percent of the wealth, income, or whatever variable is being studied. For instance, if the Lorenz curve representing income in a particular town crossed the point 0.57, 0.23 we would know that 0.57 of the population commanded just 0.23 of the town's income. In a completely equal situation, of course, 0.57 of the population would earn 0.57 of the total income, and the Lorenz curve would be identical to the 45 degree 1/1 line.

Both x and y axes are from 0 to 1, which can be expressed as a percentile (1 to 100 %) as shown in the above graph. The axes can also show quartiles.



Graphing a Lorenz Curve

To graph a Lorenz curve, the response variable (usually income or wealth) is first indexed in either equal or increasing order. Then points are graphed for a continuous distribution. If *n* is the number of instances of the response variable, then the ith x-coordinate will be i/n. The y-coordinate will be where Y_K is the response variables.

$$\frac{\sum_{k=1}^{i} Y_k}{\sum_{k=1}^{n} Y_k}$$

6.4 Check Your Progress

There are some activities to check your progress. Fill in the blanks:

(a) The variance is just the ……of the standard deviation value.

(b) Range is difference between the two.....

(c) Coefficient of variation is a …….measure of dispersion.

(d) Lorenz curve was developed by ……. in 1905.

(e) The outstanding advantage of the average deviation is its relative……..

6.5 Summary

The value given by a measure of central tendency is considered to be the representative of the whole data. It does not give the spread or range over which the data are scattered. Measures of dispersion are used to indicate this spread and the manner in which data are scattered. Commonly used measures of dispersion are: Range, Quartile deviation, Mean deviation and Standard deviation. Range is the difference between the two extreme items, i.e. it is the difference between the maximum value and minimum value in a series. Quartile deviation is obtained by dividing the difference between the upper quartile and the lower quartile by 2. Mean deviation is the mean of deviations of the items from an average (mean, median or mode). Standard Deviation is the square root of the quotient obtained by dividing the sum of squares of deviations of items from the Arithmetic mean by the number of observations. The variance is just the square of the standard deviation value. If two series differ in their units of measurement; their variability cannot be compared by any measure given so far. Hence in situations where either the two series have different units of measurements, or their means differ sufficiently in size, the coefficient of variation should be used as a measure of dispersion. It is a unit less measure of dispersion and also takes into account the size of the means of the two series. It is the best measure to compare the variability of two series or set of observations. A series with less coefficient of variation is considered more consistent. Coefficient of variation of a series of variate values is the ratio of the standard deviation to the mean multiplied by 100. A Lorenz curve is a graph used in economics to show **inequality in income spread or wealth**.

6.6 Keywords

Range: It is the difference between the maximum value and minimum value in a series.

Quartile deviation: It is obtained by dividing the difference between the upper quartile and the lower quartile by 2.

Mean deviation: It is the mean of deviations of the items from an average (mean, median or mode).

Standard Deviation: It is the square root of the quotient obtained by dividing the sum of squares of deviations of items from the Arithmetic mean by the number of observations.

Variance: It is just the square of the standard deviation value.

Coefficient of variation: It is the ratio of the standard deviation to the mean multiplied by 100.

Lorenz Curve: A Lorenz curve is a graph used in economics to show **inequality in income spread or wealth**.

6.7 Self-Assessment Tests

Q1. What does dispersion indicates about the data? Why is this of great importance?

Q2. What is the requirements of a good measure of dispersion?

Q3. Define and discuss the followingterms.

- Quartile Deviation
- Mean Deviation
- Variance
- Coefficient of Variation

Q4. Calculate mean deviation from median as well as arithmetic mean.

Class-intervals	2-4	4-6	6-8	8-10
Frequencies	3	4	2	1

Q5. Calculate the standarddeviation

Age	50-55	45-50	40-45	35-40	30-35	25-30	20-25

No.	25	30	40	45	80	110	170

Q6. Which of the two students was more consistent?

x	58	59	60	54	65	66	52	75	69	52
y	84	56	92	65	86	78	44	54	78	68

Q7. Calculate the appropriate measure ofdispersion.

Wages in rupees	No. of wage earners
Less than 35	14
35-37	62
38-40	99
41-43	18
Over 43	7

Q8. In a certain distribution with n = 25 on measurements it was found that X=56 and o=2. After these results were computed it was discovered that a mistake had been made in one of the measurements which was recorded as 64. Find the mean and standard deviation if the incorrect value 64 is omitted.

Q9. The following table gives the frequency distribution of marks obtained by students of two classes. Find the arithmetic mean, the standard deviation and coefficient of variation of the marks of two classes. Interpret the results.

Range of Marks		5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
Class	A	1	10	20	8	6	3	1	0
Class	B	5	6	15	10	5	4	2	2

6.8 Answers to Check Your Progress

(a) Square

(b) Extreme items

(c) Unit less

(d) Max Lorenz

(e) Simplicity

6.9 References/Suggested Readings

1. Gupta, S. P.: Statistical Methods, Sultan Chand and Sons, New Delhi.
2. Levin, R. I. and David, S. R.: Statistics for Management, Prentice Hall, New Delhi.
3. Gupta, C. B.: Introduction to Statistical Methods.

Subject : **Business Statistics-1**	
Course Code : **BCOM 302**	Author: Prof. M.C. Garg
Lesson No. : 7	
Measure of Skewness & Kurtosis	

STRUCTURE

7.0 LEARNING OBJECTIVES

After reading this lesson, you must be able to-

- Understand the concept, test and various measures of Skewness;
- Explain Kurtosis and Moments; and
- Make difference between skewness and dispersion.

7.1 INTRODUCTION

The measures of central tendency give us a single value that is representative of the items of a data set. The measures of dispersion give us an idea of the spread or variation of the observation about the central tendency of the items. But the measures of central tendency and dispersion do not indicate whether the distribution is symmetric or not. We may come across frequency distributions which differ widely in their nature and composition. Even then, they may have same central tendencies and dispersions. When the items in series are dispersed about the central value in even fashion, the frequency curve representing the distribution will be symmetrical. We can draw a graph with the help of given frequency distribution. If the shape of the curve, or histogram, is equal on either side of the median, it is clear that the distribution is symmetrical. If we fold the curve or histogram on the ordinate at mean, the two halves will coincide. It means the distribution is symmetrical.

7.2 CONCEPT AND TEST OF SKEWNESS

Skewness means lack of symmetry or asymmetry of a frequency distribution. Symmetry of a distribution means that the highest frequency decreases on its either side at the uniform rate and form a balanced pattern. In this case, mean, median and coincide. Consequently, skewness is said to be absent. But asymmetry or skewness is that feature of a statistical distribution which indicates that on both sides of the highest frequency mark, the frequencies do not decrease at the uniform rate. The rate of decrease on one side is more or less as compared to the other side. In such type of distribution, mean, median and mode tend to differ from each other. The larger the skewness in a series, the larger the difference between its mean and mode. Thus skewness tells us whether the spread of items from an average is symmetrical or asymmetrical.

In short, skewness is defined as opposite of symmetry and its presence tell us that the values of mean, median and mode are not equal and quartiles are not equi-distant from median. A distribution which is not symmetrical is called a skewed distribution and in such distributions, the Mean, the Median and the Mode will not coincide, but the values are

pulled apart. If the curve has a longer tail towards the right, it is said to be positively skewed. If the curve has a longer tail towards the left, it is said to be negatively skewed.

Test of Skewness: The absence of asymmetry or skewness can be stated under the following conditions. In other words, when a distribution is symmetric, the following conditions are satisfied:

1. The value of the Mean, the Median and the Mode coincide. (The values are equal).
2. Q_3- Median = Median- Q_1
3. The sum of positive deviations is equal to the sum of negative deviations.
4. The frequencies on either side of the mode are equal.
5. If plotted on a graph paper and folded at the centre of the curve(ordinate),the two halves are equal. (bell shaped curve).

	Dispersion	Skewness
1.	It shows us the spread of individual values about the central value.	1. It shows us departure from symmetry.
2.	It is useful to study the variability in data.	2. It is useful to study the concentration in in lower or higher variables.
3.	It judges the truthfulness of the central tendency.	3. It judges the differences between the central tendencies.
4.	It is a type of averages of deviation –average of the second order.	4. It is not an average, but is measured by the use of the mean, the median and the mode.
5.	It shows the degree of variability.	5. It shows whether the concentration is in higher or lower values.

MEASURES OF SKEWNESS

The measures of skewness help us the find out the direction and extent of asymmetry in a given frequency distribution. There are two measuresof skewness:

Absolute Measure

When Skewness of the series is expressed in terms of the original unit of the series, it is called absolute measure of skewness. Absolute measure of skewness tell us the extent of

asymmetry and whether it is positive or negative. It can be known by taking the difference between mean and mode. Symbolically:

$$\text{Absolute Skewness} = \overline{X} - \text{Mode}$$

If X > Z the skewness will be positive and if X < Z, the skewness will be negative. Larger the difference between mean and mode, bigger will be extent of skewness and vice-versa.

The absolute measure of skewness will not be the proper measure for comparison, and hence in each series a relative measure or coefficient of skewness will have to be computed.

Objective of Skewness : Following are the objectives of skewness :

1. Measures of skewness tell us the direction and extent of asymmetry in a series, and permit us the compare two or more series with regard to these.
2. Measures of skewness give an idea about the nature of variation of the items about the central value.

Relative Measure

The relative measure of skewness expresses the asymmetry of data in terms of some relative value of percentage. Relative measures of skewness are used for comparing the asymmetry of two or more distributions. Relative measure of skewness is known as co-efficient of Skewness. There are three important measures of relative skewness:

1. Karl Pearson's Coefficient of Skewness

According to Karl Pearson, absolute skewness = Mean - Mode. This measure is not suitable for making valid comparison of the skewness in two or more distributions, because (a) the unit of measurement may be different in different series, and (b) the same size of skewness has different significance with small or large variation in two series. Therefore, to avoid the difficulties, an absolute measure is adopted. This is done by dividing the difference between the Mean and the Mode by the Standard Deviation. The resultant coefficient is called Pearson coefficient of skewness. Thus:

$$\text{Co-efficient of Skewness } (Sk_p) = \frac{\bar{X} - \text{Mode}}{\sigma}$$

In case the mode is ill-defined, the coefficient can be determined by the changed formula :

$$\text{Coefficient of Skewness } (Sk_p) = \frac{3(\text{Mean}-\text{Median})}{\sigma}$$

$$= \frac{3(\bar{X}-M)}{\sigma}$$

llustration 1 : Calculate Karl Pearson's coefficient of skewness for thefollowing data :

25 15 23 40 27 25 23 25 20

Solution:

Computation of Mean and Standard Deviation

Size	Deviation from A= 25 (d)	Square of Deviations (d^2)
25	0	0
15	-10	100
23	-2	4
40	+15	225
27	+2	4
25	0	0
23	-2	4
25	0	0
20	-5	25
	$\Sigma d = -2$	$\Sigma d^2 = 362$

$$\text{Mean} = A \pm \frac{\Sigma d}{N}$$

$$= 25 + \frac{-2}{9}$$

$$= 25 - 0.22$$

$$= 24.78$$

$$\text{S.D.} = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

$$= \sqrt{\frac{362}{9} - \left(\frac{-2}{9}\right)^2}$$

$$= \sqrt{40.22 - (0.22)^2}$$

$$= \sqrt{40.17} = 6.3$$

Mode = 25

Karl Pearson's coefficient of skewness :

$$\frac{\text{Mean - Mode}}{\text{S.D.}}$$

$$= \frac{24.78-25}{6.3} = \frac{-0.22}{6.3} = -0.03$$

Illustration 2 : Calculate coefficient of skewness from the following:

Marks : Above	0	10	20	30	40	50	60	70	80
No. of students :	150	140	100	80	80	70	30	14	0

Solution:

Calculation of Coefficient of Skewness

Marks(X)	M.V. (m)	Frequency	Deviations from A=ÖÖ d¹	fd¹	Squares of deviation (d¹)²	(fd¹)²
0-10	5	10	-4	-40	16	160
10-20	15	40	-3	-120	9	360
20-30	25	20	-2	-40	4	80
30-40	35	0	-1	0	1	0
40-50	45	10	0	0	0	0
50-60	55	40	1	40	1	40
60-70	65	16	2	32	4	64
70-80	75	14	3	42	9	126
		N=150		Σfd¹=-86		Σ(fd¹)²=830

Mean

$$\bar{X} = 45 - \frac{86}{150} \times 10$$

$$= 45 - 5.73$$

$$= 39.27$$

Standard deviation

$$= \sqrt{\frac{\Sigma(fd^1)^2}{N} - \left(\frac{\Sigma fd^1}{N}\right)^2} \times C$$

$$= \sqrt{\frac{830}{150} - \left(\frac{-86}{150}\right)^2} \times 10$$

$$= \sqrt{5.533 - 0.328} \times 10$$

$$= \sqrt{5.205} \times 10 = 2.281 \times 10$$

$$= 22.81$$

Mode is ill-defined, therefore we must use median to calculate the coefficient of skewness :

$$\text{Median} = \frac{N}{2} = \frac{150}{2} = 75$$

Thus median lies in 40-50 group and by estimation we get :

$$M = 40 + \frac{75 - 70}{10} \times 10 = 40 + 5 = 45$$

Coefficient of Skewness

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

$$= \frac{3(39.27-45)}{22.81}$$

$$= \frac{3\ (-5.73)}{22.81}$$

$$= \frac{-17.19}{22.81}$$

$$= -0.75$$

2. Bowley's Coefficient of Skewness

In the above method of measuring skewness, the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the mean; i.e., Median - Q_1 = Q_3 - Median. This means, the value of the median is the mean of Q_1 and Q_2. But in a skewed distribution, the quartiles will not be equidistant from the Median. Hence Bowley has suggested the following formula:

$$\text{Absolute Skewness} = (Q_3 - \text{Median}) - (\text{Median} - Q_1)$$

$$= Q_3 + Q_1 - 2\ \text{Median}$$

$$\text{Coefficient of Skewness } (Sk_b) = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Illustration 3: From the information given below calculate Karl Pearson'scoefficient of skewness and also quartile coefficient of skewness:

Measure	Place A	Place B
Mean	256.5	240.8
Median	201.0	201.6
S.D.	215.4	181.1
Third quartile	260.0	242.0

First Quartile	157.0	164.2

Solution:

$$\text{Place A} \quad : \quad Sk = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

In the question mode is not given; but can be ascertained in the following method :

Mode = 3 Median – 2 Mean

Mode = 3(201) - 2(256.5)

= 603 - 513 = 90

Place B :

Since mode is not given, we have to find out :

Mode = 3 Median – 2 Mean

= (3×201.6) - (2×240.8)

= 604.8 - 481.6 = 123.2

Place A : Skewness :

$$= \frac{256.5 - 90.0}{215.4} = \frac{166.5}{215.4} = +0.773$$

Place B : Skewness :

$$= \frac{240.8 - 123.2}{181.1} = \frac{117.6}{181.1} = 0.65$$

Quartile Coefficient of Skewness

$$\text{Place A} \quad : \quad Sk = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

$$Sk = \frac{260+157 - 2\times201}{260 - 157}$$

$$= \frac{417-402}{103} = \frac{15}{103} = 0.146$$

Place B :

$$Sk = \frac{242+164.2-2(201.6)}{242 - 164.2}$$

$$= \frac{406.2 - 403.2}{77.8} = \frac{3}{77.8} = 0.0385$$

3. Kelly's Coefficient of Skewness

Bowley measures neglects the two extreme quartiles. To measure the skewness, it would be better to consider the entire data or the more extreme items. His measure of skewness is based on two deciles (i.e. 1st and 9th) andtwo percentiles (10th and 90th). Symbolically:

$$Sk_k = P_{90} + P_{10} - 2P_{50}$$

Also $$S_k = D_9 + D_1 - 2\text{Median}$$

Coefficient of skewness is defined as

$$\text{Coefficient of } S_k = \frac{P_{90}+P_{10}-2P_{50}}{P_{90} - P_{10}}$$

$$\text{Or Coefficient of } S_k = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

Illustration 4: Find out the Kelly's co-efficient of skewness of the data givenbelow:

Class	:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	:	3	10	17	7	6	4	2	1

Solution:

Class	Frequency (f)	Cumulative Frequencies (c.f.)
0-10	3	3
10-20	10	13
20-30	17	30
30-40	7	37
40-50	6	43
50-60	4	47
60-70	2	49
70-80	1	50
	Σf=50	

P_{10} = Size of $\frac{10N}{100}$ th item = $\frac{10(50)}{100}$ th item = 5 th item

P_{10} = Size of 5th item which lies in 10-20 group

$$P_{10} = l_1 + \frac{\left(\frac{10N}{100} - \text{c.f.}\right)}{f} \times i$$

$$P_{10} = 10 + \frac{(5-3)}{10} \times 10 = 10+2 = 12$$

$$P_{50} = \text{size of } \frac{50N}{100} \text{ th item} = \frac{50(50)}{100} \text{ th item} = 25 \text{ th item}$$

$$P_{50} = \text{size of 25 th item which lies in 20 – 30 group}$$

$$P_{50} = l_1 + \frac{\left(\frac{50N}{100} - \text{c.f.}\right)}{f} \times i$$

$$P_{50} = 20 + \frac{(25-13)}{17} \times 10 = 20 + \frac{120}{17} = 27.06$$

$$P_{90} = \text{size of } \left(\frac{90N}{100}\right) \text{ th item}$$

$$P_{90} = \text{size of } \left(\frac{90 \times 50}{100}\right) \text{ th item}$$

$$P_{90} = \text{size of 45 th item which lies in 50 - 60 group}$$

$$P_{90} = \frac{l_1 + 90N}{100} - \frac{\text{c.f.}}{f} \times i$$

$$P_{90} = 50 + \frac{(45-43)}{4} \times 10 = 50 + \frac{20}{4} = 55$$

$$\text{Skewness} = P_{90} + P_{10} - 2P_{50}$$

$$= 55 + 12 - 2\,(27.06) = 67 - 54.12 = 12.88$$

$$\text{Coefficient of Skewness} = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$$

$$= \frac{55 + 12 - 2(27.06)}{55 - 12} = \frac{12.88}{43} = +0.2995$$

DIFFERENCES BETWEEN SKEWNESS AND DISPERSION

The following are the main differences between skewness and dispersion:

1. **Spread of Items:** Dispersion tells us the spread of the values of items of a series. Whereas skewness informs us the direction of spread i.e. whether the spread is positive or negative.
2. **Difference Regarding Study:** Dispersion studies variation of the items of a series from its central value or average. Whereas skewness studies the asymmetry of the items from central value.
3. **Composition and Shape of Distribution:** With the help of dispersion we come to know about the composition of the distribution while skewness tells us the shape of distribution.

4. **Usefulness:** Dispersion is useful in calculation the extent of variability of the items whereas skewness tells us whether the concentration is higher in larger items or in lower items.
5. **Average:** Dispersion is the average of the deviations of items of a series from its central value. Therefore, it should be regarded as one of the measuresof central tendency. On the other hand, skewness is not an average but it is calculated by the use of averages.
6. **Measures:** The measures of dispersions are based only on the averages of the second order whereas measures of skewness are based on both averagesof first and second orders.
7. **Representation:** Dispersion tells us that upto what extent an average is a representative of a series whereas skewness tells up about the symmetry of the distribution.
8. **Variability and Asymmetry:** Dispersion generally studies the variability of a distribution whereas skewness studies the asymmetry of distribution on both sides of mode.
9. **Moments:** The measures of dispersion are based on first, second and third moments whereas the measures of skewness are based only on first and third moments.

In short, despite of above differences both dispersion and skewness are complementary to each other. When we are interested to know the extent of spread of items, we use the measures of dispersion, and when we want to know about the distribution, we take the shelter of the measures of skewness.

7.3 KURTOSIS

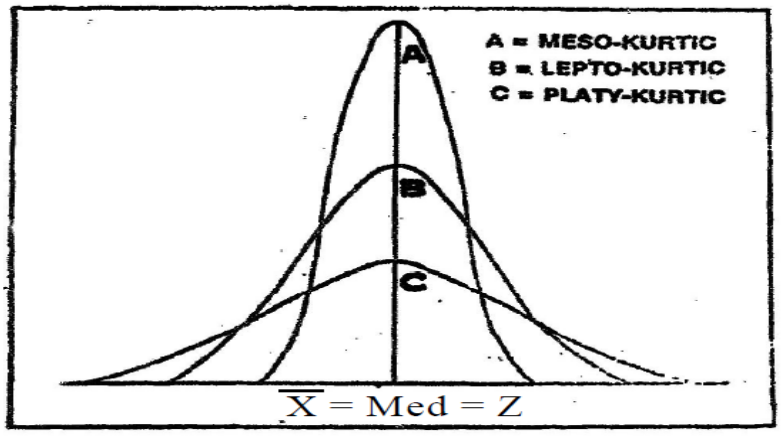
Kurtosis measures the degree of peakedness of a distribution. According to Prof. D.N. Elhance, " Another measure to test, how near a particular frequency distribution conforms to the normal curve is Kurtosis. It indicates whether a distribution is more flat-topped or more peaked than the normal distribution".

It is commonly experienced that if different distributions – even with the same mean are plotted on the graph paper, then the nature and peak form of all these curves may or may not be of the same form. In general, we find differences in the peakedness of the curves. This tendency is known as Kurtosis.

Statisticians are of the opinion that there are three distinct types of Kurtosis. These types

of Kurtosis depend on the structure and magnitude ofthe frequency distributions and also on the peakedness of the curves.

A curve having a flat top and short tails is called *Platy-Kurtic*. On the other hand, a curve with a sharp peak and long tails is called *Lepto-Kurtic*. A normal curve is smooth, continuous, perfectly symmetrical bell-shaped curve and is called *Meso-Kurtic*.



A = Meso-Kurtic
B = Lepto-Kurtic
C = Platy-Kurtic

Measures of Kurtosis

The most important measure of Kurtosis is the value of the coefficient β_2 It is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where μ_4 = 4 th moment and μ_2 = 2nd moment.

The greater the value of β_2, the more peaked at distribution.

For a normal curve the value of β_2=3.

When the value of $\beta_2$isgreater than 3, (β_2>3) the curve is more peaked than the normal curve, i.e, Lepto Kurtic.

When the value of $\beta_2$is less than 3, (β_2< 3) the curve is less peaked than the normal curve, i.e. Platy Kurtic.

The normal curve and other curves with $\beta_2 > 3$ are called Meso Kurtic.

Sometimes c_2, the derivative of β_2 is used as a measure of kurtosis. c_2, is defined as

$$C_2 = \beta_2 - 3.$$

For a normal distribution $c_2 = 0$. If c_2 is positive, the curve is Lepto Kurtic and if c_2 is negative, the curve is Platy Kurtic.

Whenever we are required to deal with the specific nature and structure of the distribution curves, we may have to deal with different types of kurtosis. In the study of statistical methods, however, kurtosis plays a less important and insignificant role.

7.4 MOMENTS

A study of Moments is of importance in studying the deviation ofa series from the normal. We have seen earlier that Mean Deviation is also known as the First Moment of Dispersion. It is the sum of the deviations of the items of a series from the mean of the series, divided by the total number of the items in the distribution. In other words, it is the average deviation of the items from the mean. The arithmetic means of the various powers of the deviations of items in a distribution from the arithmetic average of that distribution are called the Moments of the distribution. The first Moment, designated by μ_1 is therefore expressed by the formula μ_1 is, therefore, expressed by the formula

$$\mu_1 = \frac{\Sigma d}{n} \text{ or } \frac{\Sigma fd}{n}$$

because it is the mean of then first power of the deviations. Similarly, the mean of the squares of the deviations gives us the second moment about the mean. Thus, the Second, the Third, the Fourth and the nth Moments are respectively expressed as:

$$\mu_2 = \Sigma d^2/n \quad \text{or} \quad \Sigma fd^2/n$$
$$\mu_3 = \Sigma d^3/n \quad \text{or} \quad \Sigma fd^3/n$$
$$\mu_4 = \Sigma d^4/n \quad \text{or} \quad \Sigma fd^4/n$$
$$\mu_n = \Sigma d^n/n \quad \text{or} \quad \Sigma fd^n/n$$

If the value of 'd' is calculated from the arbitrary or assumed arithmetic mean, adjustments will have to be made before arriving at the various values of μ.

7.5 CHECK YOUR PROGRESS

Fill in the Blank

1. -------- indicates the difference which the items in a series individually bear to some other item taken as a standard.

2. ---------- can be defined as the arithmetic mean of various powersof deviations taken from the mean of a distribution.

3. A normal curve which is symmetrical and bell-shaped, is designated as ---------- .

4. ---------- of skewness tells us the extent of asymmetry and whether it is positive or negative.

5. ---------- measures the degree of peaked ness of a distribution.

7.6 SUMMARY

Dispersion shows the scatteredness of items in a series while skewness relates to the properties of its shape. Dispersion indicates the difference which the items in a series individually bear to some other item taken as a standard. It does not show the manner in which the items are clustered round the 'type' selected as the standard item. Skewness, on the other hand, is the measure of difference between two types of a series. It shows the manner in which the items are clustered round the 'type' by indicating the way in which items are pulled away, skewed, or distorted and render the form of the curve asymmetrical. In a symmetrical distribution the items show a perfect balance on either side of the mode. In a

skew distribution the balance is thrown to one side. The amount by which the balance exceeds on one side measures the skewness of the series.

The measures of skewness are derived from the fact that in a distribution which is perfectly symmetrical or askew, the mean, the median and the mode coincide. In a skew distribution these measurements tend to differ from each other, and the larger the skewness in a series the larger the difference between these three measures of central tendency. The difference between the mean, the median or the mode provides an easy way of expressing the lack of symmetry in a series.

Moments can be defined as the arithmetic mean of various powers of deviations taken from the mean of a distribution. Measures of Kurtosis tells up the extent to which a distribution is more peaked or more that topped than the normal curve. A normal curve which is symmetrical and bell-shaped, is designated as, Meso Kurtic. It a curve is relatively more narrow and peaked at the top, it is designated as Lepto Kurtic. If the frequency curve is more flat than normal curve, it is designated as Platy Kurtic.

7.7 KEYWORDS

- **Kurtosis** refers to the degree of flatness or peakedness in the region around the mode of a frequency curve.
- **Leptokurtic** refers to a frequency curve that is more peaked that the normal curve.
- **Measure of skewness** is the statistical technique to indicate the direction and extent of skewness in the distribution of numerical values in the data set.
- **Moments** represent a convenient and unifying method for summarizing certain descriptive statistical measures such as central tendency, variation, skewness, and kurtosis.
- **Mesokurtic** refers to a frequency curve that is a normal (symmetrical) curve.
- **Platykurtic** refers to a frequency curve that is flat-topped than the normalcurve.

7.8 SELF ASSESSMENT TESTS

1. What do you understand by skewness? What are the various methods of measuring skewness?

2. Explain the concept of skewness. Draw the sketch of a skewed frequency distribution and show the approximate position of the mean, median and mode. Give reasons why they will have the indicated position.

3. Explain briefly how the measures of skewness and kurtosis can be used in describing a frequency distribution.

4. From the following table calculate the standard deviation and coefficientof skewness:

Weekly wages in Rs.	15	20	25	30	35	40	45
No. of earners	3	25	19	16	4	5	6

5. The following table gives the data relating to marks obtained by the students appearing for the MBA Examination at the Hisar Centre: Calculate Karl Pearson's Coefficient of Skewness from the said data.

Marks group:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of students:	10	40	20	0	10	40	16	14

6. From the following data compute quartile deviation and the Coefficient of Skewness.

Size	5-7	8-10	11-13	14-16	17-19
Frequency	14	24	38	20	4

7. From a moderately skewed distribution of retail prices for men's shoes, it is found that the mean price is Rs. 20 and the median price is Rs. 17.If the coefficient of variation is 20%, find the Pearson coefficient of skewness of the distribution.

Calculate Bowley's measure of skewness from the following data:

Payment of Commission	No. of Salesmen
1000-1200	4
1200-1400	10
1400-1600	16
1600-1800	29
1800-2000	52
2000-2200	80
2200-2400	32
2400-2600	23
2600-2800	17
2800-3000	7

7.9 ANSWER TO CHECK YOUR PROGRESS

1. Dispersion
2. Moments
3. Meso Kurtic
4. Absolute measure
5. Kurtosis

7.10 REFERENCES/SUGGESTED READINGS

- Hooda RP: Statistics for Business and Economics, MacMilan India Ltd., NewDelhi, 3rd Edition, 2003.
- Aczel and Sounder Pandian: Business Sttistics, TMH, New Delhi, Fifth edition.
- Chambers JM; WS, Cleveland: Graphic Methods for Data Analysis, Boston Duxbury

Press, 1983.

- Tukey, JW: Exploratory Data Analysis, Addison Wesley Publishing, 1977.
- Gupta, SP: Statistical Methods, Sultan Chand and Sons, New Delhi, 2001.

Subject: Business Statistics-1	
Course code: BCOM 302	**Author: Dr. Anil Kumar**
Lesson: 8	**Vetter : Prof. Harbhajan Bansal**

CORRELATION

STRUCTURE

8.0 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the bivariate linear correlation
- Understand the importance of correlation analysis
- Understand the limitations of correlation analysis

8.1 INTRODUCTION

if we have information on more than one variables, we might be interested in seeing if there is any connection - any association - between them.

Statistical methods of measures of central tendency, dispersion, skewness and kurtosis are helpful for the purpose of comparison and analysis of distributions involving only one variable *i.e.* univariate distributions. However, describing the relationship between two or more variables, is another important part of statistics.

In many business research situations, the key to decision making lies in understanding the relationships between two or more variables. *For example,* in an effort to predict the behavior of the bond market, a broker might find it useful to know whether the interest rate of bonds is related to the prime interestrate. While studying the effect of advertising on sales, an account executive may find it useful to know whether there is a strong relationship between advertising dollars and sales dollars for a company.

The statistical methods of *Correlation* (discussed in the present lesson) and *Regression* (to be discussed in the next lesson) are helpful in knowing the relationship between two or more variables which may be related in same way, *like* interest rate of bonds and prime interest rate; advertising expenditure and sales; income and consumption; crop-yield and fertilizer used; height and weights and so on.

In all these cases involving two or more variables, we may be interested in seeing:

- if there is any association between the variables;
- if there is an association, is it strong enough to be useful;
- if so, what form the relationship between the two variables takes;
- how we can make use of that relationship for predictive purposes, that is, forecasting; and
- how good such predictions will be.

Since these issues are inter related, correlation and regression analysis, as two sides of a single process, consists of methods of examining the relationship between two or more variables. If two (or more) variables are correlated, we can use information about one (or more) variable(s) to predict the value of the other variable(s), and can measure the error of

estimations - *a job of regression analysis.*

8.2 WHAT IS CORRELATION?

Correlation is a measure of association between two or more variables. When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in theother variable(s), they are said to be correlated.

—The correlation between variables is a measure of the nature and degree of association between the variables‖.

As *a measure of the degree of relatedness of two variables,* correlation is widely used in exploratory research when the objective is to locate variables that might be related in some way to the variable of interest.

TYPES OF CORRELATION

Correlation can be classified in several ways. The important ways of classifying correlation are:

(i) Positive and negative,

(ii) Linear and non-linear (curvilinear) and

(iii) Simple, partial and multiple.

Positive and Negative Correlation

If both the variables move in the same direction, we say that there is a positive correlation, *i.e.,* if one variable increases, the other variable also increases on an average or if one variable decreases, the other variable also decreases on an average.

On the other hand, if the variables are varying in opposite direction, we say that it is a case of negative correlation; *e.g.,* movements of demand and supply.

Linear and Non-linear (Curvilinear) Correlation

If the change in one variable is accompanied by change in another variable in a constant

ratio, it is acase of linear correlation. Observe the following data:

X	:	10	20	30	40	50
Y	:	25	50	75	100	125

The ratio of change in the above example is the same. It is, thus, a case of linear correlation. If we plot these variables on graph paper, all the points will fall on the same straight line.

On the other hand, if the amount of change in one variable does not follow a constant ratio with the change in another variable, it is a case of non-linear or curvilinear correlation. If a couple of figures in either series *X* or series *Y* are changed, it would give a non-linear correlation.

Simple, Partial and Multiple Correlation

The distinction amongst these three types of correlation depends upon the number of variables involved in a study. If only two variables are involved in a study, then the correlation is said to be simple correlation. When three or more variables are involved in a study, then it is a problem of either partial or multiple correlation. In multiple correlation, three or more variables are studied simultaneously. But in partial correlation we consider only two variables influencing each other while the effect of other variable(s) is held constant.

Suppose we have a problem comprising three variables *X, Y* and *Z. X* is the number of hours studied, *Y* is I.Q. and *Z* is the number of marks obtained in the examination. In a multiple correlation, we will study the relationship between the marks obtained *(Z)* and the two variables, number of hours studied

(X) and I.Q. (*Y*). In contrast, when we study the relationship between *X* and *Z*, keeping an average I.Q.

(Y) as constant, it is said to be a study involving partial correlation. In this lesson, we will study linear correlation between two variables.

CORRELATION DOES NOT NECESSARILY MEAN CAUSATION

The correlation analysis, in discovering the nature and degree of relationship between variables, doesnot necessarily imply any cause and effect relationship between the variables. Two variables may be related to each other but this does not mean that one variable causes the other. *For example,* we may find that logical reasoning and creativity are correlated, but that does not mean if we could increase peoples‘ logical reasoning ability, we would produce greater creativity. We need to conduct an actual experiment to unequivocally demonstrate a causal relationship. But if it is true that influencing someones‘ logical reasoning ability does influence their creativity, then the two variables must be correlated with each other. In other words, *causation always implies correlation, however converse is not true.* Let us see some situations:

1. The correlation may be due to chance particularly when the data pertain to a small sample. A small sample bivariate series may show the relationship but such a relationship may not exist in the universe.
2. It is possible that both the variables are influenced by one or more other variables. For example, expenditure on food and entertainment for a given number of households show a positive relationship because both have increased over time. But, this is due to rise in family incomes over the same period. In other words, the two variables have been influenced by another variable - increase in family incomes.
3. There may be another situation where both the variables may be influencing each other so that we cannot say which is the cause and which is the effect. *For example,* take the case of price and demand. The rise in price of a commodity may lead to a decline in the demand for it. Here, price is the cause and the demand is the effect. In yet another situation, an increase in demand may lead to a

   rise in price. Here, the demand is the cause while price is the effect, which is just the reverse of the earlier situation. In such situations, it is difficult to identify which variable is causing the effect on which variable, as both are influencing each other.

The foregoing discussion clearly shows that correlation does not indicate any causation or functional relationship. Correlation coefficient is merely a mathematical relationship and this has nothing to do with cause and effect relation. It only reveals co-variation between two variables. Even when there is nocause-and-effect relationship in bivariate series and one interprets the relationship as causal, such a correlation is called spurious or non-sense correlation. Obviously, this will be misleading. As such, one has to be very careful in correlation exercises and look into other relevant factors before concluding a cause-and-effect relationship.

8.2.1 CORRELATION ANALYSIS

Correlation Analysis is a statistical technique used to indicate the nature and degree of relationship existing between one variable and the other(s). It is also used along with regression analysis to measure how well the regression line explains the variations of the dependent variable with the independent variable.

The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include:

1. Scatter Diagram
2. Correlation Graph
3. Pearson‘s Coefficient of Correlation
4. Spearman‘s Rank Correlation
5. Concurrent Deviation Method

8.2.2 SCATTER DIAGRAM

This method is also known as Dotogram or Dot diagram. Scatter diagram is one of the simplest methods of diagrammatic representation of a bivariate distribution. Under this method, both the variables are plotted on the graph paper by putting dots. The diagram so obtained is called "Scatter Diagram". By studying diagram, we can have rough idea about

the nature and degree of between two variables. The term scatter refers to the spreading of dots on the graph. We should keep the following points in mind while interpreting correlation:

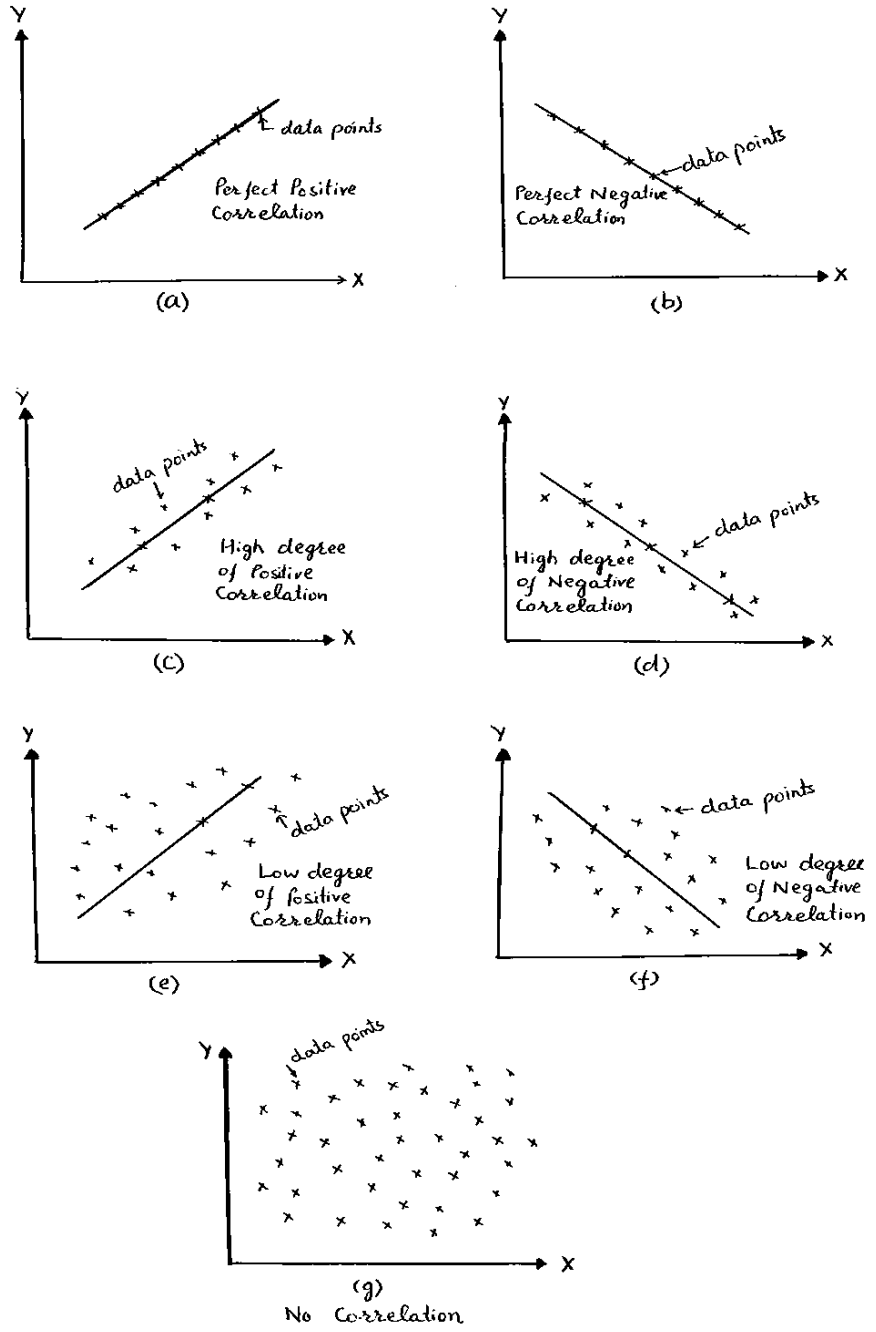


Figure 12.1 Scatter Diagrams

- if the plotted points are very close to each other, it indicates high degree of correlation. If the plotted points are away from each other, it indicates low degree of correlation.
- if the points on the diagram reveal any trend (either upward or downward), the variables are said tobe correlated and if no trend is revealed, the variables are uncorrelated.
- if there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner, the correlation is positive since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from the upper left handcorner to the lower right hand corner, the correlation is negative since in this case the values of the two variables move in the opposite directions.
- in particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points like on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative.

The various diagrams of the scattered data in Figure 8.1 depict different forms of correlation.

Example 8-1

Given the following data on sales (in thousand units) and expenses (in thousand rupees) of a firm for 10month:

Month :	J	F	M	A	M	J	J	A	S	O
Sales:	50	50	55	60	62	65	68	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

8.2.2.1 Make a Scatter Diagram

8.2.2.2 Do you think that there is a correlation between sales and expenses of the firm? Is itpositive or negative? Is it high or low?

Solution:(a) The Scatter Diagram of the given data is shown in Figure 4-2

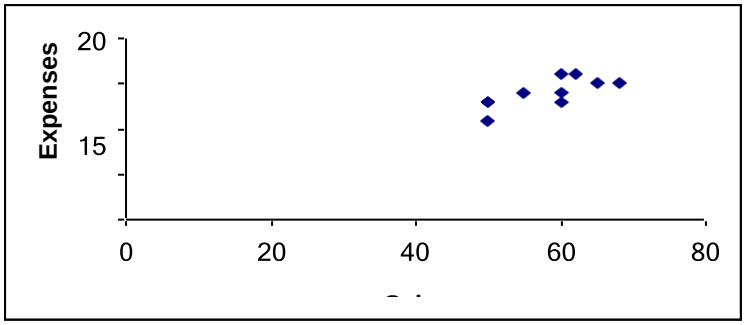


Figure 12.2 Scatter Diagram

(b) Figure 8.2 shows that the plotted points are close to each other and reveal an upward trend. So there is a high degree of positive correlation between sales and expenses of the firm.

8.2.3 CORRELATION GRAPH

This method, also known as Correlogram is very simple. The data pertaining to two series are plotted ona graph sheet. We can find out the correlation by examining the direction and closeness of two curves. If both the curves drawn on the graph are moving in the same direction, it is a case of positive correlation. On the other hand, if both the curves are moving in opposite direction, correlation is said to be negative. If the graph does not show any definite pattern on account of erratic fluctuations in the curves, then it shows an absence of correlation.

Example 8.2

Find out graphically, if there is any correlation between price yield per plot (qtls); denoted by *Y* andquantity of fertilizer used (kg); denote by *X*.

Plot No.:	1	2	3	4	5	6	7	8	9	10
Y:	3.5	4.3	5.2	5.8	6.4	7.3	7.2	7.5	7.8	8.3
X:	6	8	9	12	10	15	17	20	18	24

Solution: The Correlogram of the given data is shown in Figure 4-3

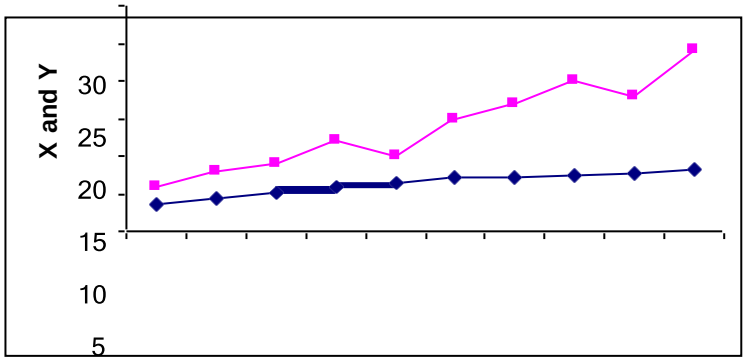


Figure 8.3 Correlation Graph

Figure 8.3 shows that the two curves move in the same direction and, moreover, they are very close to each other, suggesting a close relationship between price yield per plot (qtls) and quantity of fertilizer used (kg)

Remark: Both the Graphic methods - scatter diagram and correlation graph provide a *_feel for'* of the data – by providing visual representation of the association between the variables. These are readily comprehensible and enable us to form a fairly good, though rough idea of the nature and degree of the relationship between the two variables. However, these methods are unable to quantify the relationship between them. To quantify the extent of correlation, we make use of algebraic methods - which calculate correlation coefficient.

8.2.4 PEARSON'S COEFFICIENT OF CORRELATION

A mathematical method for measuring the intensity or the magnitude of *linear relationship* between two variables was suggested by Karl Pearson (1867-1936), a great British Biometrician and Statistician and, it is by far the most widely used method in practice.

Karl Pearson's measure, known as Pearsonian correlation coefficient between two variables *X* and *Y*, usually denoted by *r(X,Y)* or r_{xy} or simply *r* is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between *X* and *Y*, to the product of the standard deviationsof *X* and *Y*. Symbolically

$$r_{xy} = \frac{Cov(X,Y)}{S_x . S_y} \qquad \ldots\ldots\ldots\ldots(4.1)$$

when, $(X_1,Y_1);(X_2,Y_2);$ .....(X_n,Y_n) are N pairs of observations of the variables X and Y in a bivariate distribution,

$$Cov(X,Y)=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{N} \qquad \ldots\ldots\ldots(4.2a)$$

$$S_x=\sqrt{\frac{\sum(X-\bar{X})^2}{N}} \qquad \ldots\ldots\ldots(4.2b)$$

$$\text{and } S_y=\sqrt{\frac{\sum(Y-\bar{Y})^2}{N}} \qquad \ldots\ldots\ldots(4.2c)$$

Thus by substituting *Eqs. (4.2)* in *Eq. (4.1),* we can write the Pearsonian correlation coefficient as

$$r_{xy}=\frac{\frac{1}{N}\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{N}\sum(X-\bar{X})^2}\sqrt{\frac{1}{N}\sum(Y-\bar{Y})^2}}$$

$$r_{xy}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}} \qquad \ldots\ldots\ldots(4.3)$$

If we denote, $d_x=X-\bar{X}$ and $d_y=Y-\bar{Y}$

Then $r_{xy}=\frac{\sum d_x d_y}{\sqrt{\sum d_x^{\,2} d_y^{\,2}}}$ $\ldots\ldots\ldots(4.3a)$

We can further simply the calculations of *Eqs. (4.2)*

We have

$$Cov(X,Y)=\frac{1}{N}\sum(X-\bar{X})(Y-\bar{Y})$$

$$=\frac{1}{N}\sum XY-\bar{X}\bar{Y}$$

$$=\frac{1}{N}\sum XY-\frac{\sum X}{N}\frac{\sum Y}{N}$$

$$=\frac{1}{N^2}\left[N\sum XY-\sum X\sum Y\right] \qquad \ldots\ldots\ldots(4.4)$$

and $$S_x^2 = \frac{1}{N}\sum (X-\bar{X})^2$$

$$= \frac{1}{N}\sum X^2 - (\bar{X})^2$$

$$= \frac{1}{N}\sum X^2 - \left(\frac{\sum X}{N}\right)^2$$

$$= \frac{1}{N^2}\left[N\sum X^2 - (\sum X)^2\right] \qquad \ldots\ldots\ldots\ldots (4.5a)$$

Similarly, we have

$$S_y^2 = \frac{1}{N^2}\left[N\sum Y^2 - (\sum Y)^2\right] \qquad \ldots\ldots\ldots\ldots (4.5b)$$

So Pearsonian correlation coefficient may be found as

$$r_{xy} = \frac{\frac{1}{N^2}\left[N\sum XY - \sum X\sum Y\right]}{\sqrt{\frac{1}{N^2}\left[N\sum X^2 - (\sum X)^2\right]}\sqrt{\frac{1}{N^2}\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

or

$$r_{xy} = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} \qquad \ldots\ldots\ldots\ldots (4.6)$$

Remark: *Eq. (4.3)* or *Eq. (4.3a)* is quite convenient to apply if the means $\bar{X}$ and $\bar{Y}$ come out to be integers. If $\bar{X}$ or/and $\bar{Y}$ is (are) fractional then the *Eq. (4.3)* or *Eq. (4.3a)* is quite cumbersome to apply, since the computations of $\sum (X-\bar{X})^2$, $\sum (Y-\bar{Y})^2$ and $\sum (X-\bar{X})(Y-\bar{Y})$ are quite time consuming and tedious. In such a case *Eq. (4.6)* may be used provided the values of X or/ and Y are small. But if X and Y assume large values, the calculation of $\sum X^2$, $\sum Y^2$ and $\sum XY$ is again quite time consuming. Thus if *(i)* X and Y are fractional and *(ii)* X and Y assume large values, the *Eq. (4.3)* and *Eq. (4.6)* arenot generally used for numerical problems. In such cases, the step deviation method where we take the deviations of the variables X and Y from any arbitrary points is used. We will discuss this method in the properties of correlation coefficient.

Properties of Pearson Correlation Coefficient

The following are important properties of Pearson correlation coefficient:

1. *Pearson correlation coefficient cannot exceed 1 numerically*. In other words it lies between –1 and

   +1. Symbolically, $-1 \leq r \leq 1$

Remarks:

(i) This property provides us a check on our calculations. If in any problem, the obtained value of

r lies outside the limits $\pm$ 1, this implies that there is some mistake in our calculations.

(ii) The sign of *r* indicate the nature of the correlation. Positive value of *r* indicates positivecorrelation, whereas negative value indicates negative correlation. *r = 0* indicate absence of correlation.

(iii) The following table sums up the degrees of correlation corresponding to various values of *r:*

Value of *r*	Degree of correlation
±1	Perfect correlation
±0.90 or more	Very high degree of correlation
±0.75 to ±0.90	Sufficiently high degree of correlation
±0.60 to ±0.75	Moderate degree of correlation
±0.30 to ±0.60	Only the possibility of a correlation
less than ±0.30	Possibly no correlation
0	Absence of correlation

2. *Pearsonian Correlation coefficient is independent of the change of origin and scale*. Mathematically, if given variables *X* and *Y* are transformed to new variables *U* and *V* by change of origin and scale, *i e*

$$U = \frac{X - A}{h} \quad \text{and} \quad V = \frac{Y - B}{k}$$

Where *A, B, h* and *k* are constants and *h > 0, k > 0;* then the correlation coefficient between *X* and *Y* is same as the correlation coefficient between *U* and *V i.e.,*

$$r(X,Y) = r(U, V) \Rightarrow r_{xy} = r_{uv}$$

Remark: This is one of the very important properties of the correlation coefficient and is extremelyhelpful in numerical computation of *r*. We had already stated that *Eq. (4.3)* and *Eq.(4.6)* become quite tedious to use in numerical problems if *X* and/or *Y* are in fractions or if *X* and *Y* are large. In such cases we can conveniently change the origin and scale (if possible) in *X* or/and *Y* to get new variables *U* and *V*and compute the correlation between *U* and *V* by the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \sum U \sum V}{\sqrt{N\sum U^2 - (\sum U)^2}\sqrt{N\sum V^2 - (V)^2}} \quad \ldots\ldots\ldots\ldots (4.7)$$

3. Two independent variables are uncorrelated but the converse is not true

If *X* and *Y* are independent variables then $r_{xy} = 0$

However, the converse of the theorem is not true *i.e.*, uncorrelated variables need not necessarily be independent. As an illustration consider the following bivariate distribution.

X	:	1	2	3	-3	-2	-1
Y	:	1	4	9	9	4	1

For this distribution, value of *r* will be 0.

Hence in the above example the variable *X* and *Y* are uncorrelated. But if we examine the data carefully we find that *X* and *Y* are not independent but are connected by the relation $Y = X^2$. The above example illustrates that uncorrelated variables need not be independent.

Remarks: One should not be confused with the words uncorrelation and independence. $r_{xy} = 0$ *i.e.*, uncorrelation between the variables *X* and *Y* simply implies the absence of

any linear (straight line) relationship between them. They may, however, be related in some other form other than straight line *e.g.*, quadratic (as we have seen in the above example), logarithmic or trigonometric form.

4. *Pearson coefficient of correlation is the geometric mean of the two regression coefficients, i.e.*

$$r_{xy} = \pm\sqrt{b_{xy}.b_{yx}}$$

The signs of both the regression coefficients are the same, and so the value of *r* will also have the samesign. This property will be dealt with in detail in the next lesson on Regression Analysis.

5. *The square of Pearsonian correlation coefficient is known as the coefficient of determination*. Coefficient of determination, which measures the percentage variation in the dependent variable that isaccounted for by the independent variable, is a much better and useful measure for interpreting thevalue of *r*. This property will also be dealt with in detail in the next lesson.

Probable Error of Correlation Coefficient

The correlation coefficient establishes the relationship of the two variables. After ascertaining this level of relationship, we may be interested to find the extent upto which this coefficient is dependable. Probable error of the correlation coefficient is such a measure of testing the reliability of the observed value of the correlation coefficient, when we consider it as satisfying the conditions of the random sampling.

If *r* is the observed value of the correlation coefficient in a sample of *N* pairs of observations for the twovariables under consideration, then the Probable Error, denoted by *PE* (*r*) is expressed as

$$PE(r) = 0.6745\ SE(r)$$

or

$$PE(r) = 0.6745\frac{1-r^2}{\sqrt{N}}$$

There are two main functions of probable error:

1. ***Determination of limits***: The limits of population correlation coefficient are $r \pm PE(r)$, implying that if we take another random sample of the size N from the same population, then the observed value of the correlation coefficient in the second sample can be expected to lie within the limits given above, with 0.5 probability. When sample size N is small, the concept or value of PE may lead to wrong conclusions. Hence to use the concept of PE effectively, sample size N it should be fairly large.
2. ***Interpretation of 'r':*** The interpretation of 'r' based on PE is as under:
   - If $r < PE(r)$, there is no evidence of correlation, *i.e.* a case of insignificant correlation.
   - If $r > 6\ PE(r)$, correlation is significant. *If* $r < 6\ PE(r)$, it is insignificant.
   - If the probable error is small, correlation exist where $r > 0.5$

Example 8-3: Find the Pearsonian correlation coefficient between sales (in thousand units) and expenses (in thousandrupees) of the following 10 firms:

Firm:	1	2	3	4	5	6	7	8	9	10
Sales:	50	50	55	60	65	65	65	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

Solution: Let sales of a firm be denoted by X and expenses be denoted by Y

Calculations for Coefficient of Correlation

{Using *Eq. (4.3)* or *(4.3a)*}

Firm	X	Y	$d_x = X - \bar{X}$	$d_y = Y - \bar{Y}$	d_x^2	d_y^2	$d_x.d_y$
1	50	11	-8	-3	64	9	24
2	50	13	-8	-1	64	1	8
3	55	14	-3	0	9	0	0
4	60	16	2	2	4	4	4
5	65	16	7	2	49	4	14
6	65	15	7	1	49	1	7
7	65	15	7	1	49	1	7
8	60	14	2	0	4	0	0
9	60	13	2	-1	4	1	-2
10	50	13	-8	-1	64	1	8

	$\sum X$ = 580	$\sum Y$ = 140		$\sum d_x^2$ =360	$\sum d_y^2$ =22	$\sum d_x d_y$ =70

$$\overline{X} = \frac{\sum X}{N} = \frac{580}{10} = 58 \qquad \text{and} \qquad \overline{Y} = \frac{\sum Y}{N} = \frac{140}{10} = 14$$

Applying the *Eq. (4.3a),* we have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 d_y^2}}$$

$$r_{xy} = \frac{70}{\sqrt{360x22}}$$

$$r_{xy} = \frac{70}{\sqrt{7920}} = 0.78$$

The value of $r_{xy} = 0.78$, indicate a high degree of positive correlation between sales and expenses.

Example 8-4: The data on price and quantity purchased relating to a commodity for 5 months is given below:

Month :	January	February	March	April	May
Prices(Rs):	10	10	11	12	12
Quantity(Kg):	5	6	4	3	3

Find the Pearsonian correlation coefficient between prices and quantity and comment on its sign andmagnitude.

Solution: Let price of the commodity be denoted by *X* and quantity be denoted by *Y*

Calculations for Coefficient of Correlation

{Using *Eq. (4.6)*}

Month	X	Y	X^2	Y^2	XY
1	10	5	100	25	50
2	10	6	100	36	60
3	11	4	121	16	44
4	12	3	144	9	36
5	12	3	144	9	36
	$\sum X = 55$	$\sum Y = 21$	$\sum X^2 = 609$	$\sum Y^2 = 95$	$\sum XY = 226$

Applying the *Eq. (4.6), we* have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r_{xy} = \frac{5x226 - 55x21}{\sqrt{(5x609 - 55x55)(5x95 - 21x21)}}$$

$$r_{xy} = \frac{1130 - 1155}{\sqrt{20x34}} =$$

$$r_{xy} = \frac{-25}{\sqrt{680}}$$

$$r_{xy} = -0.98$$

The negative sign of *r* indicate negative correlation and its large magnitude indicate a very high degreeof correlation. So there is a high degree of negative correlation between prices and quantity demanded.

Example 8-5

Find the Pearsonian correlation coefficient from the following series of marks obtained by 10 studentsin a class test in mathematics (*X)* and in Statistics (*Y)*:

X : 45 70 65 30 90 40 50 75 85 60

Y : 35 90 70 40 95 40 60 80 80 50

Also calculate the Probable Error.

Solution:

Calculations for Coefficient of Correlation

{Using *Eq. (4.7)*}

X	Y	U	V	U^2	V^2	UV
45	35	-3	-6	9	36	18
70	90	2	5	4	25	10
65	70	1	1	1	1	1
30	40	-6	-5	36	25	30
90	95	6	6	36	36	36
40	40	-4	-5	16	25	20
50	60	-2	-1	4	1	2
75	80	3	3	9	9	9
85	80	5	3	25	9	15
60	50	0	-3	0	9	0
		$\sum U = 2$	$\sum V = -2$	$\sum U^2 = 140$	$\sum V^2 = 176$	$\sum UV = 141$

We have, defined variables U and V as

$$U = \frac{X-60}{5} \qquad \text{and} \qquad V = \frac{Y-65}{5}$$

Applying the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \left(\sum U \sum V\right)}{\sqrt{N\sum U^2 - \left(\sum U\right)^2}\sqrt{N\sum V^2 - \left(\sum V\right)^2}}$$

$$= \frac{10x141 - 2x(-2)}{\sqrt{10x140 - 2x2}\sqrt{10x176 - (-2)x(-2)}}$$

$$= \frac{1410+4}{\sqrt{1400-4}\sqrt{1760-4}}$$

$$= \frac{1414}{\sqrt{2451376}} = 0.9$$

So, there is a high degree of positive correlation between marks obtained in Mathematics and in Statistics.

Probable Error, denoted by $PE\ (r)$ is given as

$$PE(r) = 0.6745\frac{1-r^2}{\sqrt{N}}$$

$$PE(r) = 0.6745\frac{1-(0.9)^2}{\sqrt{10}}$$

$$PE(r) = 0.0405$$

So the value of r is highly significant.

8.2.5 SPEARMAN'S RANK CORRELATION

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, *etc.*, which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904, which consists in obtaining the correlation coefficient between the

ranks of *N* individuals in the two attributes under study.

Suppose we want to find if two characteristics *A*, say, intelligence and *B*, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of *N* individuals in order of merit (ranks) *w.r.t.* proficiency in the two characteristics. Let the random variables *X* and *Y* denote the ranks of the individuals in the characteristics *A* and *B* respectively. If we assume that there is no tie, *i.e.,* if no two individuals get the same rank in a characteristic then, obviously, *X* and *Y* assume numerical values ranging from *1* to *N*.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for the group of individuals.

Spearman's rank correlation coefficient, usually denoted by ρ(Rho) is given by the equation

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} \qquad \ldots\ldots\ldots\ldots (4.8)$$

Where *d* is the difference between the pair of ranks of the same individual in the two characteristics and *N* is the number of pairs.

Example 8-6

Ten entries are submitted for a competition. Three judges study each entry and list the ten in rank order.Their rankings are as follows:

Entry:	A	B	C	D	E	F	G	H	I	J
Judge J_1:	9	3	7	5	1	6	2	4	10	8
Judge J_2:	9	1	10	4	3	8	5	2	7	6
Judge J_3:	6	3	8	7	2	4	1	5	9	10

Calculate the appropriate rank correlation to help you answer the following questions:

(i) Which pair of judges agrees the most?

(ii) Which pair of judges disagrees the most?

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq.(4.8)*}

Entry	Rank by Judges			Difference in Ranks					
	J_1	J_2	J_3	$d(J_1\&J_2)$	d^2	$d(J_1\&J_3)$	d^2	$d(J_2\&J_3)$	d^2
A	9	9	6	0	0	+3	9	+3	9
B	3	1	3	+2	4	0	0	-2	4
C	7	10	8	-3	9	-1	1	+2	4
D	5	4	7	+1	1	-2	4	-3	9
E	1	3	2	-2	4	-1	1	+1	1
F	6	8	4	-2	4	+2	4	+4	16
G	2	5	1	-3	9	+1	1	+4	16
H	4	2	5	+2	4	-1	1	-3	9
I	10	7	9	+3	9	+1	1	-2	4
J	8	6	10	+2	4	-2	4	-4	16
					$\sum d^2 = 48$		$\sum d^2 = 26$		$\sum d^2 = 88$

$$\rho(J_1 \& J_2) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 48}{10(10^2-1)} = 1 - \frac{288}{990} = 1 - 0.29 = +0.71$$

$$\rho(J_1 \& J_3) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 26}{10(10^2-1)} = 1 - \frac{156}{990} = 1 - 0.1575 = +0.8425$$

$$\rho(J_2 \& J_3) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 88}{10(10^2-1)} = 1 - \frac{528}{990} = 1 - 0.53 = +0.47$$

So *(i)* Judges J_1 and J_3 agree the most

(ii) Judges J_2 and J_3 disagree the most

Spearman's rank correlation *Eq.(4.8)* can also be used even if we are dealing with variables, which are measured quantitatively, *i.e.* when the actual data but not the ranks relating to two variables are given. In such a case we shall have to convert the data into ranks. The highest (or the smallest) observation is given the rank 1. The next highest (or the next lowest) observation is given rank 2 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for all the variables under consideration.

Example 8-7

Calculate the rank coefficient of correlation from the following data:

X:	75	88	95	70	60	80	81	50
Y:	120	134	150	115	110	140	142	100

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq.(4.8)*}

X	Ranks R_X	Y	Ranks R_Y	$d = R_X - R_Y$	d^2
75	5	120	5	0	0
88	2	134	4	-2	4
95	1	150	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
80	4	140	3	+1	1
81	3	142	2	+1	1
50	8	100	8	0	0

$\sum d^2 = 6$

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - \frac{6 \times 6}{8(8^2-1)} = 1 - \frac{36}{504} = 1 - 0.07 = +0.93$$

Hence, there is a high degree of positive correlation between *X* and *Y*

Repeated Ranks

In case of attributes if there is a tie *i.e.*, if any two or more individuals are placed together in any classification *w.r.t.* an attribute or if in case of variable data there is more than one item with the same value in either or both the series then Spearman's *Eq.(4.8)* for calculating the rank correlation coefficient breaks down, since in this case the variables *X* [the ranks of individuals in characteristic A (1^{st} series)] and *Y* [the ranks of individuals in characteristic B (2^{nd} series)] do not take the values from *1* to *N*.

In this case common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks, which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common rank. For example, suppose an item is repeated at rank 4. Then the common rank to be assigned to each item is (4+5)/2, *i.e.*, 4.5 which is the average of 4 and 5, the ranks which these observations would have assumed if they were different.The next item will be assigned the rank

6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will be (7+8+9)/3, *i.e.*, 8 which is the arithmetic mean of 7,8 and 9 *viz.*, the ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

If only a small proportion of the ranks are tied, this technique may be applied together with *Eq.(4.8)*. If a large proportion of ranks are tied, it is advisable to apply an adjustment or a correction factor to *Eq.(4.8)*as explained below:

—In the Eq.(4.8) add the factor

$$\frac{m(m^2-1)}{12} \qquad \text{.....(4.8a)}$$

to $\sum d^2$*; where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the series*‖.

Example 8-8

For a certain joint stock company, the prices of preference shares (X) and debentures (Y) are givenbelow:

X:	73.2	85.8	78.9	75.8	77.2	81.2	83.8
Y:	97.8	99.2	98.8	98.3	98.3	96.7	97.1

Use the method of rank correlation to determine the relationship between preference prices anddebentures prices.

Solution: *Calculations for Coefficient of Rank Correlation*

{Using *Eq. (4.8)* and *(4.8a)*}

X	Y	Rank of X (X_R)	Rank of Y (Y_R)	$d = X_R - Y_R$	d^2
73.2	97.8	7	5	2	4
85.8	99.2	1	1	0	0
78.9	98.8	4	2	2	4
75.8	98.3	6	3.5	2.5	6.25
77.2	98.3	5	3.5	1.5	2.25
81.2	96.7	3	7	-4	16
83.8	97.1	2	6	-4	16
				$\sum d = 0$	$\sum d^2 = 48.50$

In this case, due to repeated values of Y, we have to apply ranking as average of 2 ranks, which could have been allotted, if they were different values. Thus ranks 3 and 4 have been allotted as 3.5 to both the values of Y = 98.3. Now we also have to apply correction factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m in the number of times the value is repeated, here $m = 2$.

$$\rho = \frac{6\left[\sum d^2 + \frac{m(m^2-1)}{2}\right]}{N(N^2-1)} = \frac{6\left[48.5 + \frac{2(4-1)}{12}\right]}{7(7^2-1)} = 1 - \frac{6 \times 49}{7 \times 48} = 0.125$$

Hence, there is a very low degree of positive correlation, probably no correlation, between preference share prices and debenture prices.

Remarks on Spearman's Rank Correlation Coefficient

1. We always have $\sum d = 0$, which provides a check for numerical calculations.
2. Since Spearman's rank correlation coefficient, ρ, is nothing but Karl Pearson's correlation coefficient, *r*, between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.
3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure, which is distribution free (or non-parametric). Spearman's ρ is such a distribution free measure, since no strict assumption are made about the from of the population from which sample observations are drawn.
4. Spearman's formula is easy to understand and apply as compared to Karl Pearson's formula. The values obtained by the two formulae, *viz* Pearsonian *r* and Spearman's ρ are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank correlation should beonly slightly lower than the Pearsonian coefficient.
5. Spearman's formula is the only formula to be used for finding correlation coefficient if we aredealing with qualitative characteristics, which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.
6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution. For *N* >30, this formula should not be used unless the ranks are given.

8.2.6 CONCURRENT DEVIATION METHOD

This is a casual method of determining the correlation between two series when we are not very serious about its precision. This is based on the signs of the deviations(*i.e.*the direction of the change)of the values of the variable from its preceding value and does not take into account the exactmagnitude of the values of the variables. Thus we put a plus (+) sign, minus (-) sign or equality (=) sign for the deviation if the value of the variable is greater than, less than or equal to the preceding value respectively. The

deviations in the values of two variables are said to be concurrent if they have the same sign (either both deviations are positive or both are negative or both are equal). The formula used for computing correlation coefficient r_c by this method is given by

$$r_c = \pm\sqrt{\pm\left(\frac{2c-N}{N}\right)} \quad \ldots\ldots\ldots\ldots (4.9)$$

Where c is the number of pairs of concurrent deviations and N is the number of pairs of deviations. If ($2c$-N) is positive, we take positive sign in and outside the square root in *Eq. (4.9)* and if ($2c$-N) is negative, we take negative sign in and outside the square root in *Eq. (4.9).*

Remarks: (i) It should be clearly noted that here N is not the number of pairs of observations but it is the number of pairs of deviations and as such it is one less than the number of pairs of observations.

(ii) Coefficient of concurrent deviations is primarily based on the following principle: *—If the short time fluctuations of the time series are positively correlated or in other words, if their deviations are concurrent, their curves would move in the same direction and would indicate positive correlation between them‖*

Example 8-9

Calculate coefficient of correlation by the concurrent deviation method

Supply:	112	125	126	118	118	121	125	125	131	135
Price:	106	102	102	104	98	96	97	97	95	90

Solution:

Calculations for Coefficient of Concurrent Deviations

{Using *Eq. (4.9)*}

Supply (X)	Sign of deviation from preceding value (X)	Price (Y)	Sign of deviation preceding value (Y)	Concurrent deviations
112		106		
125	+	102	-	

126	+	102	=	
118	-	104	+	
118	=	98	-	
121	+	96	-	
125	+	97	+	+(*c*)
125	=	97	=	= (*c*)
131	+	95	-	
135	+	90	-	

We have

Number of pairs of deviations, N =10 – 1 = 9

c = Number of concurrent deviations

= Number of deviations having like signs

= 2

Coefficient of correlation by the method of concurrent deviations is given by:

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

$$r_c = \pm\sqrt{\pm\left(\frac{2x2-9}{9}\right)}$$

$$r_c = \pm\sqrt{\pm(-0.5556)}$$

Since $2c - N$ = -5 (negative), we take negative sign inside and outside the square root

$$r_c = -\sqrt{-(-0.5556)}$$

$$r_c = -\sqrt{0.5556}$$

$$r_c = -0.7$$

Hence there is a fairly good degree of negative correlation between supply and price.

8.2.7 LIMITATIONS OF CORRELATION ANALYSIS

As mentioned earlier, correlation analysis is a statistical tool, which should be properly used so that correct results can be obtained. Sometimes, it is indiscriminately used by management, resulting in misleading conclusions. We give below some *errors* frequently made in the use of correlation analysis:

1. Correlation analysis cannot determine cause-and-effect relationship. One should not assume that a change in *Y* variable is caused by a change in *X* variable unless one is reasonably sure that one variable is the cause while the other is the effect. Let us take an example.

   Suppose that we study the performance of students in their graduate examination and their earnings after, say, three years of their graduation. We may find that these two variables are highly and positively related. At the same time, we must not forget that both the variables might have been influenced by some other factors such as quality of teachers, economic and social status of parents, effectiveness of the interviewing process and so forth. If the data on these factors are available, thenit is worthwhile to use multiple correlation analysis instead of bivariate one.
2. Another mistake that occurs frequently is on account of misinterpretation of the coefficient of correlation. Suppose in one case $r = 0.7$, it will be wrong to interpret that correlation explains 70 percent of the total variation in *Y*. The error can be seen easily when we calculate the coefficient of determination. Here, the coefficient of determination r^2 will be 0.49. This means that only 49 percent of the total variation in *Y* is explained.

   Similarly, the coefficient of determination is misinterpreted if it is also used to indicate causal relationship, that is, the percentage of the change in one variable is due to the change in another variable.
3. Another mistake in the interpretation of the coefficient of correlation occurs when one concludes a positive or negative relationship even though the two variables are actually unrelated. For example, the age of students and their score in the examination have no relation with each other. The two variables may show similar movements but there does not seem to be a common link between them.

To sum up, one has to be extremely careful while interpreting coefficient of

correlation. Before one concludes a causal relationship, one has to consider other relevant factors that might have any influence on the dependent variable or on both the variables. Such an approach will avoid many of the pitfalls in the interpretation of the coefficient of correlation. It has been rightly said that the coefficient of correlation is not only one of the most widely used, but also one of the widely abused statistical measures.

8.3 PARTIAL CORRELATION

It is a statistical technique to analyze the association between dependent variables and one of the independent variables by eliminating the effect of other variables. Partial correlation is also known as Net Correlation. It is a statistical technique to study the association between one dependent variable and one independent variable by keeping other independent variables constant. In simple correlation, the effect of other independent variables was ignored. It is mainly divided into three categories:

- Zero order coefficient
- First order coefficient
- Second order coefficient

8.3.1 Zero Order Coefficient

It is simple correlation between two variables whereby no other variables are held constant. It can beunderstood with the help of following example:

It is assumed that there are three variables named as x_1, x_2 and x_3. In this case, simple correlation will becalculated between any of two variables by ignoring third variable completely in the calculation. Then,

- r_{12} = Simple correlation between x_1 and x_2 variables by ignoring x_3 variable.
- r_{23} = Simple correlation between x_2 and x_3 variables by ignoring x_1 variable.
- r_{31} = Simple correlation between x_3 and x_1 variables by ignoring x_2 variable.

The formulas for calculating simple correlation between two variables have already been discussed in section 8.3 of this lesson. Therefore, you can refer to the section 8.3 to understand the simple correlation calculation techniques.

8.3.2 First Order Coefficient

It is a partial correlation between two variables keeping the third variable constant. You can keep any one variable constant out of three variables in final calculation of partial correlation. Hence, there canbe three cases of first order coefficient:

- $r_{12.3}$ = Simple correlation between x_1 and x_2 variables keeping x_3 variable constant.
- $r_{23.1}$ = Simple correlation between x_2 and x_3 variables keeping x_1 variable constant.
- $r_{13.2}$ = Simple correlation between x_1 and x_3 variables keeping x_2 variable constant.Following are the formulas for calculating first order partial correlation:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)}\ \sqrt{(1 - r_{23}^2)}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{(1 - r_{12}^2)}\ \sqrt{(1 - r_{13}^2)}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1 - r_{12}^2)}\ \sqrt{(1 - r_{23}^2)}}$$

Example 8.10 The simple coefficients of correlation between two variables out of three are: r_{12} = 0.8,r_{13} = 0.7 and r_{23} = 0.6 Find the partial coefficients of correlation, $r_{12.3}$, $r_{23.1}$ and $r_{13.2}$.

Solution:

$$r_{12.3} = \frac{0.8 - (0.7 \times 0.6)}{\sqrt{(1 - 0.7^2)}\ \sqrt{(1 - 0.6^2)}} = \frac{0.8 - 0.42}{0.714 \times 0.8} = \frac{0.38}{0.5712} = 0.665$$

$$r_{23.1} = \frac{0.6 - (0.8 \times 0.7)}{\sqrt{(1 - 0.8^2)}\ \sqrt{(1 - 0.7^2)}} = \frac{0.6 - 0.56}{0.6 \times 0.714} = \frac{0.04}{0.4284} = 0.093$$

$$r_{13.2} = \frac{0.7 - (0.8 \times 0.6)}{\sqrt{(1 - 0.8^2)}\ \sqrt{(1 - 0.6^2)}} = \frac{0.7 - 0.48}{0.6 \times 0.8} = \frac{0.22}{0.48} = 0.458$$

8.3.3 Second Order Coefficient

It is a partial correlation between two variables keeping other two variables constant. When four variables are included in a problem than any two of them are kept constant and the correlation is calculated between remaining two variables. Let's assume that there are four variables including x_1, x_2, x_3 and x_4. Thus, if correlation is calculated between x_1 and x_2 then two variables x_3 and x_4 will be kept constant. It can be symbolically represented by $r_{12.34}$. In this way, there can be six cases of partial correlation of second order including:

- $r_{12.34}$ = correlation between x_1 and x_2 variables keeping x_3 and x_4 constant.
- $r_{13.24}$ = correlation between x_1 and x_3 variables keeping x_2 and x_4 constant.
- $r_{14.23}$ = correlation between x_1 and x_4 variables keeping x_2 and x_3 constant.
- $r_{23.14}$ = correlation between x_2 and x_3 variables keeping x_1 and x_4 constant.
- $r_{24.13}$ = correlation between x_2 and x_4 variables keeping x_1 and x_3 constant.
- $r_{34.12}$ = correlation between x_3 and x_4 variables keeping x_1 and

x_2 constant.Following are the formulas for calculating second order partial correlation:

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\, r_{23.4}}{\sqrt{(1 - r^2_{13.4})}\ \sqrt{(1 - r^2_{23.4})}}$$

$$r_{13.24} = \frac{r_{13.4} - r_{12.4}\, r_{23.4}}{\sqrt{(1 - r^2_{12.4})}\ \sqrt{(1 - r^2_{23.4})}}$$

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}\, r_{24.3}}{\sqrt{(1 - r^2_{12.3})}\ \sqrt{(1 - r^2_{24.3})}}$$

$$r_{23.14} = \frac{r_{23.4} - r_{12.4}\, r_{13.4}}{\sqrt{(1 - r^2_{12.4})}\ \sqrt{(1 - r^2_{13.4})}}$$

$$r_{24.13} = \frac{r_{24.3} - r_{12.3}\, r_{14.3}}{\sqrt{(1 - r^2_{12.3})}\ \sqrt{(1 - r^2_{14.3})}}$$

$$r_{34.12} = \frac{r_{34.2} - r_{13.2}\, r_{14.2}}{\sqrt{(1 - r^2_{13.2})}\ \sqrt{(1 - r^2_{14.2})}}$$

1. The partial correlation coefficient lies between -1 and +1.
2. These correlation coefficients are calculated on the basis of zero order coefficient or simple coefficients of correlation where no variable is kept constant. Thus r_{12} is a simple correlationcoefficient between x_1 and x_2.

Example: 8.11 Given: $r_{12.4} = 0.6$, $r_{13.4} = 0.5$ and $r_{23.4} = 0.7$. Find $r_{12.34}$ and $r_{13.24}$.

Solution:

$$r_{12.34} = \frac{0.6 - (0.5 \times 0.7)}{\sqrt{(1 - 0.5^2)}\ \sqrt{(1 - 0.7^2)}}$$

$$= \frac{0.6-0.35}{\sqrt{0.75}\ \sqrt{0.51}} = \frac{0.25}{0.867 \times 0.714} = \frac{0.25}{0.619} = 0.404$$

$$r_{13.24} = \frac{0.5 - (0.6 \times 0.7)}{\sqrt{(1 - 0.6^2)}\ \sqrt{(1 - 0.7^2)}}$$

$$= \frac{0.5-0.42}{\sqrt{0.64}\ \sqrt{0.51}} = \frac{0.8}{0.8 \times 0.714} = \frac{0.8}{0.5712} = 0.14$$

Example 8.12 It is observed that weight of fish depends on food consumption, type of water and temperature. Find the partial correlation between weight of fish (x_1) and temperature (x_4) keeping food consumption (x_2) and type of water (x_3) constant. Given that $r_{14.3} = 0.6$, $r_{12.3} = 0.5$, $r_{24.3}$ 0.7. Find $r_{14.23}$.

Solution:

$$r_{14.23} = \frac{0.6 - (0.5 \times 0.7)}{\sqrt{(1 - 0.5^2)}\ \sqrt{(1 - 0.7^2)}}$$

$$= \frac{0.6-0.35}{\sqrt{0.75}\ \sqrt{0.51}} = \frac{0.25}{0.867 \times 0.714} = \frac{0.25}{0.619} = 0.404$$

This gives the partial correlation between weight of fish and temperature keeping food consumption andtype of water constant.

8.3.4 Limitations of Partial Correlation

Followings are the main limitations of partial correlation:

1. In the calculation of partial correlation coefficient, it is presumed that there should exist a linearrelation between variables. In real situation, this condition lacks in some cases.
2. The independent variable in the study of partial correlation should be linearly independent.
3. The reliability of partial correlation coefficient decreases as their order goes up. This means that the second order partial coefficients are not as dependable as the first order ones are. Therefore, it is necessary that the size of the items in the gross correlation should be large.
4. It involves a lot of calculation work and its analysis is not easy.

8.4 MULTIPLE CORRELATION:

Multiple correlations is a technique to study the relationship between three or four variables simultaneously. In this, effect of all independent variables on a dependent variable is analyze. The coefficient of multiple linear correlations is represented by R. It can be understood as follow:

Assuming three variable x_1, x_2 and x_3 whereas x_1 is the dependent variable and other two are independent variables. Here, the multiple correlation coefficients can be defined as follow:

- $R_{1.23}$ = Multiple correlation coefficient with x_1 as dependent variable and x_2 and x_3 as independentvariables.
- $R_{2.13}$ = Multiple correlation coefficient with x_2 as dependent variable and x_1 and x_3 as independentvariables.
- $R_{3.12}$ = Multiple correlation coefficient with x_3 as dependent variable and x_1 and x_2 as independentvariables.

8.4.1 Calculation of Multiple Correlation Coefficient

Following are the formulas for calculating multiple correlation coefficients:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13} r_{23}}{1 - r_{23}^2}}$$

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}\, r_{13} r_{23}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}\, r_{13} r_{23}}{1 - r_{12}^2}}$$

If there are three independent variables and one dependent variable the formula for finding out the multiple correlation is

$$R_{1.234} = \sqrt{1 - (1 - r^2_{14})\,(1 - r^2_{12.3})\,(1 - r^2_{12.34})}$$

Remarks:

1. Multiple correlation coefficients is a positive coefficient always. Its value ranges between 0 and 1. Itcannot be a negative value.

2. If $R_{1.23} = 0$, then $r_{12} = 0$ and $r_{13} = 0$.

3. $R_{1.23} \geq r_{12}$ and $R_{1.23} \geq r_{13}$

4. $R_{1.23}$ is the same thing as $R_{1.32}$. The position of the subscript to the right of the dot does not make adifference.

5. By squaring $R_{1.23}$, we get the coefficient of multiple determination.

Example 8.13 The following zero order correlation coefficient are given: $r_{12} = 0.5$, $r_{13} = 0.6$, $r_{23} = 0.7$.Calculate the multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

Solution:

$$R_{1.23} = \sqrt{\frac{(0.5)^2 + (0.6)^2 - 2 \times 0.5 \times 0.6 \times 0.7}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{0.25 + 0.36 - 0.42}{0.51}} = \sqrt{\frac{0.19}{0.51}} = \sqrt{0.3725} = 0.610$$

$$R_{2.13} = \sqrt{\frac{(0.5)^2 + (0.7)^2 - 2 \times 0.5 \times 0.6 \times 0.7}{1 - (0.6)^2}}$$

$$= \sqrt{\frac{0.25 + 0.49 - 0.42}{0.64}} = \sqrt{\frac{0.32}{0.64}} = \sqrt{0.5} = 0.707$$

$$R_{3.12} = \sqrt{\frac{(0.6)^2 + (0.7)^2 - 2 \times 0.5 \times 0.6 \times 0.7}{1 - (0.5)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.42}{0.75}} = \sqrt{\frac{0.43}{0.75}} = \sqrt{0.5733} = 0.757$$

8.4.2 Limitations of Multiple Correlation

Followings are the main limitations of multiple correlation coefficients:It also assume a linear relationship between the zero order coefficients like partial correlationcoefficients.

1. It assume that the independent variables affect the dependent variable in an independent manner andhave an additive property.
2. If, however, there is interrelation between independent variables, their effect cannot be distinct norcan they be additive.
3. The calculation of multiple correlation is cumbersome and difficult.

8.5 CHECK YOUR PROGRESS

1. If both the variables move in the same direction, we say that there is a .....correlation.
2. If the change in one variable is accompanied by change in another variable in a constant ratio, it is acase of correlation.
3. When three or more variables are involved in a study, then it is a problem of either

……correlation.

4. The Pearsonian correlation coefficient between the ranks X and Y is called the……

between the characteristics A and B for the group of individuals.

5. Correlation analysis cannot determine .....relationship.

8.6 SUMMARY

The statistical methods of Correlation (discussed in the present lesson) and Regression (to be discussed in the next lesson) are helpful in knowing the relationship between two or more variables which may be related in same way, *like* interest rate of bonds and prime interest rate. Correlation can be classified in several ways. The important ways of classifying correlation are: Positive and negative, Linear and non- linear (curvilinear) and Simple, partial and multiple. The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include: Scatter Diagram, Correlation Graph, Pearson‘s Coefficient of Correlation, Spearman‘s Rank Correlation and Concurrent Deviation Method. There are measure uses of correlation analysis, but some limitations too. It has been rightly said that the coefficient of correlation is not only one of the most widely used, but also one of the widely abused statistical measures.

8.7 KEYWORDS

Correlation: When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in the other variable(s), they are said to be correlated.

Scatter Diagram: Under this method, both the variables are plotted on the graph paper by putting dots. By studying diagram, we can have rough idea about the nature and degree of relationship between two variables.

Correlation Graph: The data pertaining to two series are plotted on a graph sheet and then find out the correlation by examining the direction and closeness of two curves.

Pearson's correlation coefficient: It is denoted by $r(X,Y)$ or r_{xy} or simply r is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between X and Y, to the product of the standard deviations of X and Y.

Concurrent Deviation Method: The deviations in the values of two variables are said to be concurrentif they have the same sign.

8.8 SELF-ASSESSMENT TEST

1. ―Correlation and Regression are two sides of the same coin‖. Explain.
2. Explain the meaning and significance of the concept of correlation. Does correlation always signify casual relationships between two variables? Explain with illustration on what basis can the followingcorrelation be criticized?
   (a) Over a period of time there has been an increased financial aid to under developed countries and also an increase in comedy act television shows. The correlation is almost perfect.
   (b) The correlation between salaries of school teachers and amount of liquor sold during the period 1940 – 1980 was found to be 0.96
3. Write short not on the following
   (a) Spurious correlation
   (b) Positive and negative correlation
   (c) Linear and non-linear correlation
   (d) Simple, multiple and partial correlation
4. What is a scatter diagram? How does it help in studying correlation between two variables, in respectof both its nature and extent?
5. Write short note on the following
   (a) Karl Pearson's coefficient of correlation
   (b) Probable Error
   (c) Spearman's Rank Correlation Coefficient
   (d) Coefficient of Concurrent Deviation
6. Draw a scatter diagram from the data given below and interpret it.

X:	10	20	30	40	50	60	70	80

Y:	32	20	24	36	40	28	38	44

7. Calculate Karl Pearson's coefficient of correlation between expenditure on advertising (*X*) and sales

   (*Y*) from the data given below:

X:	39	65	62	90	82	75	25	98	36	78
Y:	47	53	58	86	62	68	60	91	51	84

8. To study the effectiveness of an advertisement a survey is conducted by calling people at random byasking the number of advertisements read or seen in a week (*X*) and the number of items purchased

   (*Y*) in that week.

X:	5	10	4	0	2	7	3	6
Y:	10	12	5	2	1	3	4	8

   Calculate the correlation coefficient and comment on the result.

9. Calculate coefficient of correlation between *X* and *Y* series from the following data and calculate itsprobable error also.

X:	78	89	96	69	59	79	68	61
Y:	125	137	156	112	107	136	123	108

10. In two set of variables *X* and *Y*, with 50 observations each, the following data are observed:

    $\overline{X}$ = 10, SD of *X* = 3
    $\overline{Y}$ = 6, SD of *Y* = 2 $r_{xy} = 0.3$

    However, on subsequent verification, it was found that one value of *X* (=10) and one value of *Y* (= 6) were inaccurate and hence weeded out with the remaining 49 pairs of values. How the original value of is $r_{xy} = 0.3$ affected?

11. Calculate coefficient of correlation *r* between the marks in statistics (*X*) and Accountancy (*Y*) of 10 students from the following:

X:	52	74	93	55	41	23	92	64	40	71
Y:	45	80	63	60	35	40	70	58	43	64

    Also determine the probable error or *r*.

12. The coefficient of correlation between two variables *X* and *Y* is 0.48. The covariance is 36. Thevariance of *X* is 16. Find the standard deviation of *Y*.

13. Twelve entries in painting competition were ranked by two judges as shown below:

Entry:	A	B	C	D	E	F	G	H	I	J
Judge I:	5	2	3	4	1	6	8	7	10	9
Judge II:	4	5	2	1	6	7	10	9	3	8

Find the coefficient of rank correlation.

14. Calculate Spearman's rank correlation coefficient between advertisement cost (*X*) and sales (*Y*)from the following data:

X:	39	65	62	90	82	75	25	98	36	78
Y:	47	53	58	86	62	68	60	91	51	84

15. An examination of eight applicants for a clerical post was taken by a firm. From the marks obtainedby the applicants in the Accountancy (*X*) and Statistics (*Y*) paper, compute rank coefficient of correlation.

Applicant:	A	B	C	D	E	F	G	H
X:	15	20	28	12	40	60	20	80
Y:	40	30	50	30	20	10	30	60

16. Calculate the coefficient of concurrent deviation from the following data:

Year:	1993	1994	1995	1996	1997	1998	1999	2000
Supply:	160	164	172	182	166	170	178	192
Price:	222	280	260	224	266	254	230	190

17. Obtain a suitable measure of correlation from the following data regarding changes in price indexof the shares *A* and *B* during nine months of a year:

Month:	A	M	J	J	A	S	O	N	D
A:	+4	+3	+2	-1	-3	+4	-5	+1	+2
B:	-2	+5	+3	-2	-1	-3	+4	-1	-3

18. The cross-classification table shows the marks obtained by 105 students in the subjects of Statistics and Finance:

		Marks in Statistics				
		50-54	55-59	60-64	65-74	Total
Marks in Finance	50-59	4	6	8	7	25
	60-69	-	10	12	13	35
	70-79	16	9	20	-	45
	80-89	-	-	-	-	-

Total	20	25	40	20	105

Find the coefficient of correlation between marks obtained in two subjects.

8.9 ANSWERS TO CHECK YOUR PROGRESS

1. Positive
2. Linear
3. Partial or multiple
4. Rank correlation coefficient
5. Cause-and-effect

8.10 REFERNCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers andDistributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., NewDelhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill PublishingCompany Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of IndiaPvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics-1	
Course Code: BCOM-302	**Author: Anil Kumar**
Lesson: 9	**Vetter: Prof. Harbhajan Bansal**

REGRESSION

STRUCTURE

9.0 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the concept of linear regression
- Understand the importance of regression analysis
- Understand the limitations of regression analysis.

9.1 WHAT IS REGRESSION?

—if we find any association between two or more variables, we might be interested in estimating the value of one variable for known value(s) of another variable(s)‖

In business, several times it becomes necessary to have some forecast so that the management can takea decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets. Thus, to determine the nature and extent of relationship between these two variables becomes important for the company.

In the preceding lesson, we studied in some depth linear correlation between two variables. Here we have a similar concern, the association between variables, except that we develop it further in two respects. *First,* we learn how to build statistical models of relationships between the variables to have a better understanding of their features. *Second,* we extend the models to consider their use in forecasting. For this purpose, we have to use the technique - regression analysis - which forms the subject-matter of this lesson.

In 1889, Sir Francis Galton, a cousin of Charles Darwin published a paper on heredity, ―*Natural Inheritance*‖. He reported his discovery that sizes of seeds of sweet pea plants appeared to ‗revert' or

‗regress', to the mean size in successive generations. He also reported results of a study of the relationship between heights of fathers and heights of their sons. A straight line was fit to the data pairs: height of father versus height of son. Here, too, he found a ‗regression to mediocrity' The heights of the sons represented a movement

away from their fathers, towards the average height. We credit Sir Galton with the idea of statistical regression.

While most applications of regression analysis may have little to do with the ―regression to the mean‖ discovered by Galton, the term ―regression‖ remains. It now refers to the statistical technique of modeling the relationship between two or more variables. In general sense, regression analysis means the estimation or prediction of the unknown value of one variable from the known value(s) of the other variable(s). It is one of the most important and widely used statistical techniques in almost all sciences - natural, social or physical.

In this lesson we will focus only on simple regression –linear regression involving only two variables: a dependent variable and an independent variable. Regression analysis for studying more than two variables at a time is known as multiple regressions.

INDEPENDENT AND DEPENDENT VARIABLES

Simple regression involves only two variables; one variable is predicted by another variable. *The variable to be predicted* is called the dependent variable. *The predictor* is called the independent variable, or *explanatory variable.* For example, when we are trying to predict the demand for television sets on the basis of population growth, we are using the demand for television sets as the dependent variable and the population growth as the independent or predictor variable.

The decision, as to which variable is which sometimes, causes problems. Often the choice is obvious, as in case of demand for television sets and population growth because it would make no sense to suggest that population growth could be dependent on TV demand! The population growth has to be the independent variable and the TV demands the dependent variable.

If we are unsure, here are some points that might be of use:

- If we have control over one of the variables then that is the independent. For example, a manufacturer can decide how much to spend on advertising and expect his sales to be dependent upon how much he spends
- It there is any lapse of time between the two variables being measured, then the latter

must depend upon the former, it cannot be the other way round

- If we want to predict the values of one variable from your knowledge of the other variable, thevariable to be predicted must be dependent on the known one

9.1.1 LINEAR REGRESSION

The task of bringing out linear relationship consists of developing methods of fitting a straight line, or a regression line as is often called, to the data on two variables.

The line of Regression is the graphical or relationship representation of the best estimate of one variablefor any given value of the other variable. The nomenclature of the line depends on the independent and dependent variables. If *X* and *Y* are two variables of which relationship is to be indicated, a line that gives best estimate of *Y* for any value of *X*, it is called *Regression line of Y on X*. If the dependent variable changes to *X*, then best estimate of *X* by any value of *Y* is called Regression line of X on Y.

9.1.2 REGRESSION LINE OF *Y* ON *X*

For purposes of illustration as to how a straight line relationship is obtained, consider the sample paired data on sales of each of the $N = 5$ months of a year and the marketing expenditure incurred in each month, as shown in Table 5-1

Table 13-1

Month	Sales (Rs lac) *Y*	Marketing Expenditure (Rs thousands) *X*
April	14	10
May	17	12
June	23	15
July	21	20
August	25	23

Let *Y*, the sales, be the dependent variable and *X*, the marketing expenditure, the independent variable. We note that for each value of independent variable *X*, there is a specific value of the dependent variable *Y*, so that each value of *X* and *Y* can be seen as paired observations.

9.1.2.1 Scatter Diagram

Before obtaining a straight-line relationship, it is necessary to discover whether the relationship betweenthe two variables is linear, that is, the one which is best explained by a straight line. A good way of doing this is to plot the data on *X* and *Y* on a graph so as to yield a scatter diagram, as may be seen in Figure 13-1. A careful reading of the scatter diagram reveals that:

- The overall tendency of the points is to move upward, so the relationship is positive
- The general course of movement of the various points on the diagram can be best explained by astraight line
- There is a high degree of correlation between the variables, as the points are very close to eachother

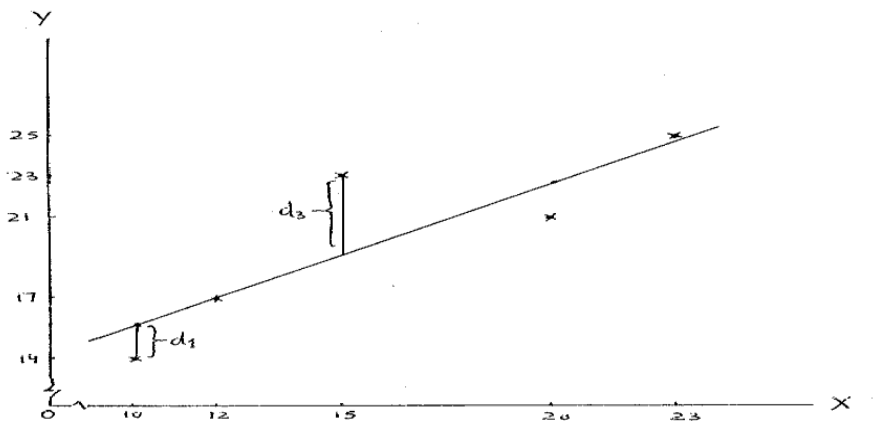


Figure 13-1 Scatter Diagram with Line of Best Fit

9.1.2.2 Fitting a Straight Line on the Scatter Diagram

If the movement of various points on the scatter diagram is best described by a straight line, the next step is to fit a straight line on the scatter diagram. It has to be so fitted that on the whole it lies as closeas possible to every point on the scatter diagram. The necessary requirement for meeting this condition being that the sum of the squares of the vertical deviations of the observed Y values from the straight line is minimum.

$$d_1^2 + d_2^2 + \ldots\ldots\ldots\ldots\ldots + d_N^2 = \sum_{j=1}^{N} d_j^2$$

As shown in Figure 5-1, if d_l, d_2,..., d_N are the vertical deviations' of observed Y values from the straight line, fitting a straight line requires that is the minimum. The deviations d_j have to be squared to avoid negative deviations canceling out the positive deviations. Since a straight line so fitted best approximates all the points on the scatter diagram, it is better known as the best approximating line or the line of best fit. A line of best fit can be fitted by means of:

1. Free hand drawing method, and
2. Least square method

Free Hand Drawing:

Free hand drawing is the simplest method of fitting a straight line. After a careful inspection of the movement and spread of various points on the scatter diagram, a straight line is drawn through these points by using a transparent ruler such that on the whole it is closest to every point. A straight line so drawn is particularly useful when future approximations of the dependent variable are promptly required.

Whereas the use of free hand drawing may yield a line nearest to the line of best fit, the major drawback is that the slope of the line so drawn varies from person to person because of the influence of subjec- tivity. Consequently, the values of the dependent variable estimated on the basis of such a line may not be as accurate and precise as those based on the line of best fit.

Least Square Method:

The least square method of fitting a line of best fit requires minimizing the sum of the squares of vertical deviations of each observed Y value from the fitted line. These deviations, such as d_1 and d_3, are shown in Figure 5-1 and are given by $Y - Y_c$, where Y is the observed value and Y_c the corresponding computed value given by the fitted line

for the i^{th} value of X.

$$Y_c = a + bX_i \quad \ldots\ldots\ldots\ldots(5.1)$$

The straight line relationship in *Eq.(5.1),* is stated in terms of two constants a and b

- The constant a is the Y-intercept; it indicates the height on the vertical axis from where the straight line originates, representing the value of Y when X is zero.
- Constant b is a measure of the slope of the straight line; it shows the absolute change in Y for a unit change in X. As the slope may be positive or negative, it indicates the nature of relationship between Y and X. Accordingly, b is also known as the regression coefficient of Y on X.

Since a straight line is completely defined by its intercept a and slope b, the task of fitting the same reduces only to the computation of the values of these two constants. Once these two values are known, the computed Y_c values against each value of X can be easily obtained by substituting X values in the linear equation.

In the method of least squares the values of a and b are obtained by solving simultaneously the following pair of normal equations

$$\sum Y = aN + b\sum X \ldots\ldots\ldots\ldots (5.2)$$

$$\sum XY = a\sum X + b\sum X^2 \quad \ldots\ldots(5.2)$$

The value of the expressions - $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ can be obtained from the given

observations and then can be substituted in the above equations to obtain the value of a and b.

Since simultaneous solving the two normal equations for a and b may quite often be cumbersome andtime consuming, the two values can be directly obtained as

$$a = \bar{Y} - b\bar{X} \quad \ldots\ldots\ldots\ldots (5.3)$$

and

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} \quad \ldots\ldots\ldots (5.4)$$

Note: *Eq. (5.3) is obtained simply by dividing both sides of the first of Eqs. (5.2) by N and Eq.(5.4) is obtained by substituting* $(Y - bX)$ *in place of a in the second of Eqs. (5.2)* Instead of directly computing *b*, we may first compute value of *a* as

$$a = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N\sum X^2 - \left(\sum X\right)^2} \quad \ldots\ldots\ldots\ldots (5.5)$$

and

$$b = \frac{\bar{Y} - a}{\bar{X}} \quad \ldots\ldots\ldots\ldots (5.6)$$

Note: *Eq. (5.5) is obtained by substituting* $\frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$ *for b in Eq. (5.3) and Eq. (5.6) is obtained simply by rearranging Eq. (5.3)*

Table 13-2

Computation of a *and* b

Y	X	XY	X^2	Y^2
14	10	140	100	196
17	12	204	144	289
23	15	345	225	529
21	20	420	400	441
25	23	575	529	625
$\sum Y = 100$	$\sum X = 80$	$\sum XY = 1684$	$\sum X^2 = 1398$	$\sum Y^2 = 2080$

So using *Eqs. (5.5)* and *(5.4)*

$$a = \frac{100x1398 - 80x1684}{5x1398 - (80)^2} = \frac{139800 - 134720}{6990 - 6400} = \frac{5080}{590} = 8.6101695$$

$$b = \frac{5x1684 - 80x100}{5x1398 - (80)^2} = \frac{8420 - 8000}{6990 - 6400} = \frac{420}{590} = 0.7118644$$

Now given $a = 8.61$ and $b = 0.71$

The regression *Eq.(5.1)* takes the form

$$Y_c = 8.61 + 0.71X \ldots\ldots\ldots\ldots (5.1a)$$

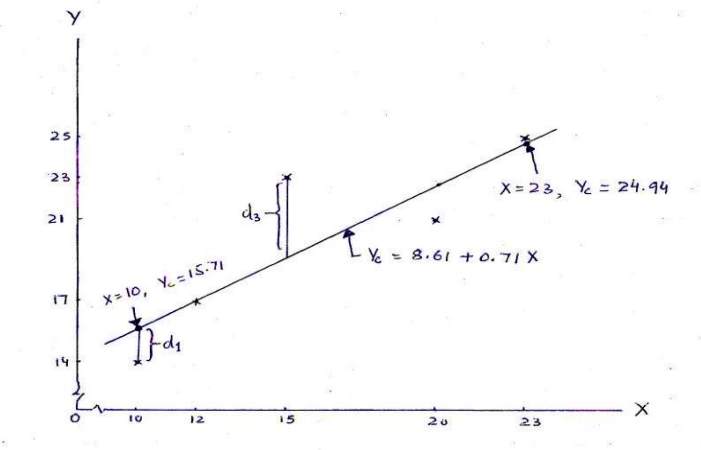


Figure 13-2 Regression Line of *Y* on *X*

Then, to fit the line of best fit on the scatter diagram, only two computed Y_c values are needed. These can be easily obtained by substituting any two values of X in *Eq. (5.1a).* When these are plotted on the diagram against their corresponding values of X, we get two points, by joining which (by means of a straight line) gives us the required line of best fit, as shown in Figure 5-2

Some Important Relationships:

We can have some important relationships for data analysis, involving other measures such as X , Y , S_x, S_y and the correlation coefficient r_{xy}.

Substituting $\overline{Y} - b\overline{X}$ [from *Eq.(5.3)*] for *a* in *Eq.(5.1)*

$$Y_c = (\overline{Y} - b\overline{X}) + bX$$

or

$$Y_c - \overline{Y} = b(X - \overline{X}) \ldots\ldots\ldots\ldots (5.7)$$

Dividing the numerator and denominator of *Eq.(5.4)* by N^2, we get

$$b = \frac{\frac{\sum XY}{N} - \left(\frac{\sum X}{N}\right)\left(\frac{\sum Y}{N}\right)}{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$$

$$\text{or } b = \frac{\frac{\sum XY}{N} - \overline{X}\overline{Y}}{S_x^2}$$

$$\text{or } b = \frac{Cov(X,Y)}{S_x^2} \qquad \ldots\ldots\ldots\ldots (5.8)$$

We know, coefficient of correlation, r_{xy} is given by

$$r_{xy} = \frac{Cov(X,Y)}{S_x S_y}$$

$$\text{or} \quad Cov(X,Y) = r_{xy} S_x S_y$$

So *Eq. (5.8)* becomes

$$b = r_{xy} \frac{S_x S_y}{S_x^2}$$

$$b = r_{xy} \frac{S_y}{S_x} \ldots\ldots\ldots\ldots (5.9)$$

Substituting $r_{xy} \frac{S_y}{S_x}$ for *b* in *Eq.(5.7)*, we get

$$Y_c - \overline{Y} = r_{xy} \frac{S_y}{S_x} (X - \overline{X}) \ldots\ldots\ldots (5.10)$$

These are important relationship for data analysis.

9.1.1.1 Predicting an Estimate and its Preciseness

The main objective of regression analysis is to know the nature of relationship between two variables and to use it for predicting the most likely value of the dependent

variable corresponding to a given, known value of the independent variable. This can be done by substituting in *Eq.(5.1a)* any known value of *X* corresponding to which the most likely estimate of *Y* is to be found.

For example, the estimate of Y (*i.e.* Y_c), corresponding to $X = 15$ is

$$Y_c = 8.61 + 0.71(15)$$
$$= 8.61 + 10.65$$
$$= 19.26$$

It may be appreciated that an estimate of *Y* derived from a regression equation will not be exactly the same as the *Y* value which may actually be observed. The difference between estimated Y_c values and the corresponding observed *Y* values will depend on the extent of scatter of various points around the line of best fit.

The closer the various paired sample points *(Y, X)* clustered around the line of best fit, the smaller the difference between the estimated Y_c and observed *Y* values, and vice-versa. On the whole, the lesser the scatter of the various points around, and the lesser the vertical distance by which these deviate from the line of best fit, the more likely it is that an estimated Y_c value is close to the corresponding observed *Y* value.

The estimated Y_c values will coincide the observed *Y* values only when all the points on the scatter diagram fall in a straight line. If this were to be so, the sales for a given marketing expenditure could have been estimated with l00 percent accuracy. But such a situation is too rare to obtain. Since some of the points must lie above and some below the straight line, perfect prediction is practically non-existent in the case of most business and economic situations.

This means that the estimated values of one variable based on the known values of the other variable are always bound to differ. The smaller the difference, the greater the precision of the estimate, and vice- versa. Accordingly, the preciseness of an estimate can be obtained only through a measure of the magnitude of error in the estimates,

called the error of estimate.

9.1.1.2 Error of Estimate

A measure of the error of estimate is given by the standard error of estimate of *Y on X*, denoted as S_{yx} and defined as

$$S_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N}} \quad \ldots\ldots\ldots\ldots (5.11)$$

S_{yx} measures the average absolute amount by which observed Y values depart from the corresponding computed Y_c values. Computation of S_{yx} becomes little cumbersome where the number of observations N is large. In suchcases S_{yx} may be computed directly by using the equation:

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b\sum XY}{N}} \quad \ldots\ldots\ldots (5.12)$$

By substituting the values of $\sum Y^2$, $\sum Y$, and $\sum XY$ from the Table 5-2, and the calculated values of a and b

We have

$$S_{yx} = \sqrt{\frac{2080 - 8.61x100 - 0.71x1684}{5}}$$

$$= \sqrt{\frac{2080 - 861 - 1195.64}{5}} = \sqrt{\frac{23.36}{5}}$$

$$= \sqrt{4.67} = 2.16$$

Interpretations of S_{yx}

A careful observation of how the standard error of estimate is computed reveals the following:

1. S_{yx} is a concept statistically parallel to the standard deviation S_y. The only difference between the two being that the standard deviation measures the dispersion around the mean; the standard error of estimate measures the dispersion around the regression line. Similar to the property of arithmetic mean, the sum of the deviations of different Y values from their corresponding estimated Y_c values is equal to zero. That is

$$\Sigma(Y_i - \bar{Y}) = \Sigma(Y_i - Y_c) = 0 \text{ where } i = 1, 2, \ldots, N.$$

2. S_{yx} tells us the amount by which the estimated Y_c values will, on an average, deviate from

the observed Y values. Hence it is an estimate of the average amount of error in the estimated Y_c values. The actual error (the residual of Y and Y_c) may, however, be smaller or larger than the average error. Theoretically, these errors follow a normal distribution. Thus, assuming that $n \geq 30$, $Y_c \pm 1.S_{yx}$ means that 68.27% of the estimates based on the regression equation will be within $1.S_{yx}$ Similarly, $Y_c \pm 2.S_{yx}$ means that 95.45% of the estimates will fall within $2.S_{yx.}$ Further, for the estimated value of sales against marketing expenditure of Rs 15 thousand being Rs 19.26 lac, one may like to know how good this estimate is. Since S_{yx} is estimated to be Rs 2.16 lac, itmeans there are about 68 chances (68.27) out of 100 that this estimate is in error by not more than Rs 2.16 lac above or below Rs 19.26 lac. That is, there are 68% chances that actual sales would fallbetween (19.26 - 2.16) = Rs 17.10 lac and (19.26 + 2.16) = Rs 21.42 lac.

3. Since S_{yx} measures the closeness of the observed Y values and the estimated Y_c values, it also serves as a measure of the reliability of the estimate. Greater the closeness between the observed and estimated values of Y, the lesser the error and, consequently, the more reliable the estimate. And vice-versa.
4. Standard error of estimate S_{yx} can also be seen as a measure of correlation insofar as it expresses the degree of closeness of scatter of observed Y values about the regression line. The closer the observed Y values scattered around the regression line, the higher the correlation between the two variables.

A major difficulty in using S_{yx} as a measure of correlation is that it is expressed in the same units of measurement as the data on the dependent variable. This creates problems in situations requiring comparison of two or more sets of data in terms of correlation. It is mainly due to this limitation that the standard error of estimate is not generally used as a measure of correlation. However, it does serve as the basis of evolving the coefficient of determination, denoted as r^2, which provides an alternate methodof obtaining a measure of correlation.

9.1.2 REGRESSION LINE OF *X* ON *Y*

So far we have considered the regression of Y on X, in the sense that Y was in the role of dependent and X in the role of an independent variable. In their reverse position,

such that X is now the dependent and Y the independent variable, we fit a line of regression of X on Y. The regression equation in this case will be

$$X_c = a' + b'Y \text{.....} (5.13)$$

Where X_c denotes the computed values of X against the corresponding values of Y. a' is the X-interceptand b' is the slope of the straight line. Two normal equations to solve $a'$and b' are

$$\sum X = a'N + b'\sum Y \text{ .....} (5.14)$$

$$\sum XY = a'\sum Y + b'\sum Y^2 \text{ .....} (5.14)$$

The value of a' and b' can also be obtained directly

$$a' = \overline{X} - b'\overline{Y} \text{.....} (5.15)$$

and

$$b' = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - \left(\sum Y\right)^2} \text{.....} (5.16)$$

or

$$a' = \frac{\sum X \sum Y^2 - \sum Y \sum XY}{N\sum Y^2 - \left(\sum Y\right)^2} \text{ .....} (5.17)$$

and

$$b' = \frac{\overline{X} - a'}{\overline{Y}} \text{ .....} (5.18)$$

$$b' = \frac{Cov(Y, X)}{S_y^2} \text{.....} (5.19)$$

$$b' = r_{yx} \frac{S_x}{S_y} \text{.....} (5.20)$$

So, Regression equation of X on Y may also be written as

$$X_c - \bar{X} = b'(Y-\bar{Y}) \quad \text{.....(5.21)}$$

$$X_c - \bar{X} = r_{yx} \frac{S_x}{S_y}(Y - \bar{Y}) \quad \text{..... (5.22)}$$

As before, once the values of a' and b' have been found, their substitution in *Eq.(5.13)* will enable us toget an estimate of X corresponding to a known value of Y

Standard Error of estimate of X on Y *i.e.* S_{xy} will be

$$S_{xy} = \sqrt{\frac{(X - X_c)^2}{N}} \quad \text{.....(5.23)}$$

or

$$S_{xy} = \sqrt{\frac{\sum X^2 - a'\sum X - b'\sum XY}{N}} \quad \text{.....(5.24)}$$

For example, if we want to estimate the marketing expenditure to achieve a sale target of Rs 40 lac, wehave to obtain regression line of X on Y *i. e.*

$$X_c = a' + b'Y$$

So using *Eqs. (5.17)* and *(5.16),* and substituting the values of $\sum X, \sum Y^2, \sum Y$ and $\sum XY$ 5-2, we have from Table

$$a' = \frac{80x2080 - 100x1684}{5x2080 - (100)^2} = \frac{166400 - 168400}{10400 - 10000} = \frac{-2000}{400} = -5.00$$

$$b' = \frac{5x1684 - 80x100}{5x2080 - (100)^2} = \frac{8420 - 8000}{10400 - 10000} = \frac{420}{400} = 1.05$$

Now given that $a' = -5.00$ and $b' = 1.05$, Regression equation *(5.13)* takes the form

$$X_c = -5.00 + 1.05Y$$

So when $Y = 40$(Rs lac), the corresponding X value is

$$X_c = -5.00 + 1.05x40$$
$$= -5 + 42$$
$$= 37$$

That is to achieve a sale target of Rs 40 lac, there is a need to spend Rs 37 thousand on marketing.

9.2 PROPERTIES OF REGRESSION COEFFICIENTS

As explained earlier, the slope of regression line is called the regression coefficient. It tells the effect on dependent variable if there is a unit change in the independent variable. Since for a paired data on X and Y variables, there are two regression lines: regression line of Y on X and regression line of X on Y, so we have two regression coefficients:

(i) Regression coefficient of Y on X, denoted by b_{yx} [b in *Eq.(5.1)*]

(ii) Regression coefficient of X on Y, denoted by b_{xy} [b' in *Eq.(5.13)*]

The following are the important properties of regression coefficients that are helpful in data analysis

1. The value of both the regression coefficients cannot be greater than 1. However, value of both the coefficients can be below 1 or at least one of them must be below 1, so that the square root of the product of two regression coefficients must lie in the limit ±1.
2. Coefficient of correlation is the geometric mean of the regression coefficients, *i.e.*

$$r = \pm\sqrt{b.b'} \quad \ldots\ldots\ldots\ldots (5.25)$$

The signs of both the regression coefficients are the same, and so the value of *r* will also have the same sign.

3. The mean of both the regression coefficients is either equal to or greater than the coefficient ofcorrelation, *i.e.*

$$\frac{b+b'}{2} \geq r$$

4. Regression coefficients are independent of change of origin but not of change of scale. Mathematically, if given variables *X* and *Y* are transformed to new variables *U* and *V* by change of origin and scale, *i. e.*

$$U = \frac{X-A}{h} \quad \text{and} \quad V = \frac{Y-B}{k}$$

Where *A, B, h* and *k* are constants, $h > 0, k > 0$ then

Regression coefficient of *Y* on *X* = *k/h* (Regression coefficient of *V* on *U*)

$$b_{yx} = \frac{k}{h} b_{vu} \quad \text{and}$$

Regression coefficient of *X* on *Y* = *h/k* (Regression coefficient of *U* on *V*)

$$b_{xy} = \frac{h}{k} b_{uv}$$

5. Coefficient of determination is the product of both the regression coefficients *i.e.*

$$r^2 = b.b'$$

9.2.1 REGRESSION LINES AND COEFFICIENT OF CORRELATION

The two regression lines indicate the nature and extent of correlation between the variables.

The two regression lines can be represented as

$$Y - \bar{Y} = r\frac{S_y}{S_x}(X - \bar{X}) \qquad \text{and} \qquad X - \bar{X} = r\frac{S_x}{S_y}(Y - \bar{Y})$$

We can write the slope of these lines, as

$$b = r\frac{S_y}{S_x} \qquad \text{and} \qquad b' = r\frac{S_x}{S_y}$$

If θ is the angle between these lines, then

$$\tan\theta = \frac{b - b'}{1 + bb'} = \frac{S_x S_y}{S_x^2 + S_y^2}\left(\frac{r^2 - 1}{r}\right)$$

$$\text{or } \theta = \tan^{-1}\left[\frac{S_x S_y}{S_x^2 + S_y^2}\left(\frac{r^2 - 1}{r}\right)\right] \quad \ldots\ldots\ldots\ldots (5.26)$$

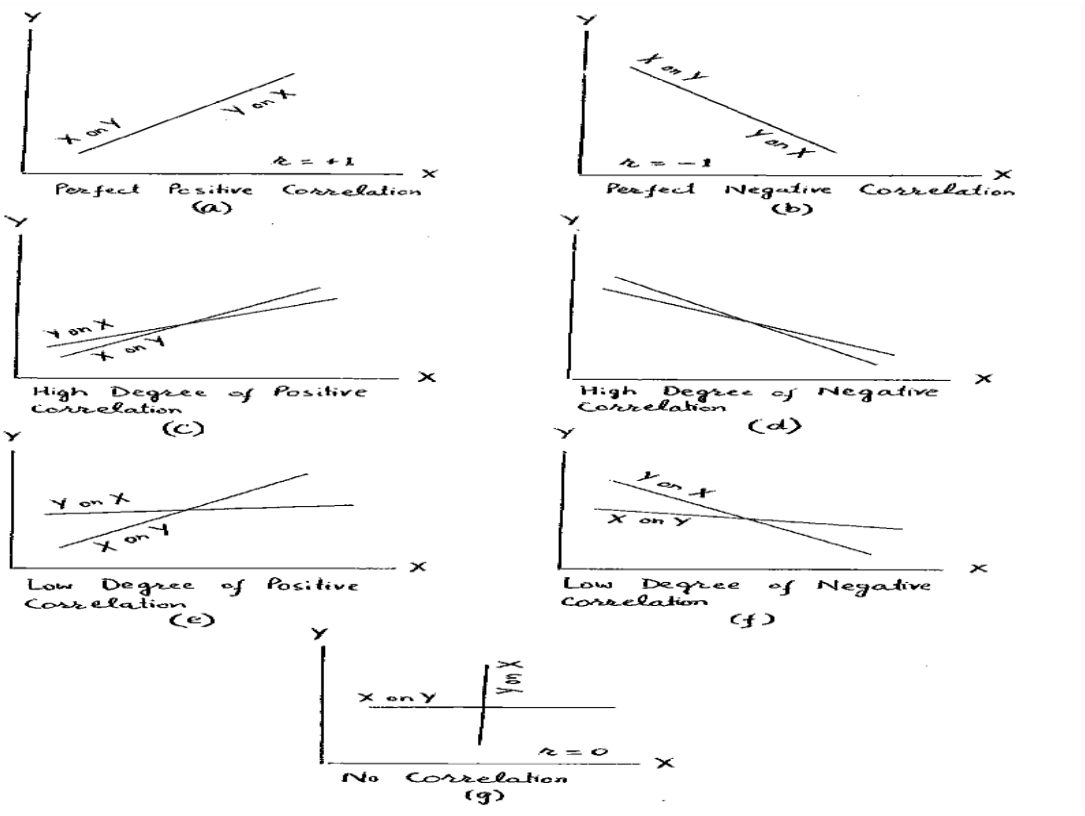


Figure 13-3 Regression Lines and Coefficient of Correlation

Eq. (5.26) **reveals the following:**

- In case of perfect positive correlation (r = +1) and in case of perfect negative correlation (r = -1), θ = 0, so the two regression lines will coincide, *i.e.* we have only one line, see (a) and (b) in Figure 5-3.

  The farther the two regression lines from each other, lesser will be the degree of correlation andnearer the two regression lines, more will be the degree of correlation, see (c) and (d) in Figure 5-3.

- If the variables are independent *i.e.* r = 0, the lines of regression will cut each other at right angle.See (g) in Figure 5-3.

Note: *Both the regression lines cut each other at mean value of X and mean value of Y i.e.* $\bar{X}$ *and* $\bar{Y}$.

9.2.2 COEFFICIENT OF DETERMINATION

Coefficient of determination gives the percentage variation in the dependent variable that is accounted for by the independent variable. In other words, the coefficient of determination gives the ratio of the explained variance to the total variance. The coefficient of determination is given by the square of the correlation coefficient, *i.e.* r^2. Thus, Coefficient of determination

$$r^2 = \frac{ExplainedVariance}{TotalVariance}$$

$$r^2 = \frac{\sum(Y_c - \bar{Y})^2}{\sum Y - \bar{Y}} \quad \ldots\ldots\ldots(5.27)$$

We can calculate another coefficient K^2, known as coefficient of Non-Determination, which is the ratio of unexplained variance to the total variance.

$$K^2 = \frac{Unexplained Variance}{TotalVariance}$$

$$K^2 = \frac{\sum(Y - Y_c)^2}{\sum(Y - \bar{Y})} \quad \ldots\ldots(5.28)$$

$$K^2 = 1 - \frac{ExplainedVariance}{TotalVariance}$$

$$= 1 - r^2 \ldots\ldots(5.29)$$

The square root of the coefficient of non-determination, *i.e.* K gives the coefficient of alienation

$$K = \pm\sqrt{1 - r^2} \quad \ldots\ldots\ldots(5.30)$$

Relation Between S_{yx} and r:

A simple algebraic operation on *Eq. (5.30)* brings out some interesting points about the relation between S_{yx} and r. Thus, since

$$\sum_c (Y - Y)^2 = N S_{yx}^2 \quad \text{and} \quad \sum (Y - \bar{Y})^2 = N S_y^2$$

So we have coefficient of Non-determination

$$K^2 = \frac{\sum_c (Y - Y)^2}{\sum (Y - \bar{Y})^2}$$

$$K^2 = \frac{N S_{yx}^2}{N S_y^2} = \frac{S_{yx}^2}{S_y^2}$$

So $$1 - r^2 = \frac{S_{yx}^2}{S_y^2}$$

or $$\frac{S_{yx}}{S_y} = \sqrt{1 - r^2} \quad \ldots\ldots\ldots\ldots (5.31)$$

If coefficient of correlation, r, is defined as the under root of the coefficient of determination

$$r = \sqrt{r^2}$$

$$r^2 = 1 - \frac{S_{yx}^2}{S_y^2}$$

$$r = \sqrt{1 - \frac{S_{yx}}{S_y^2}} \quad \ldots\ldots\ldots\ldots (5.32)$$

On carefully observing *Eq. (5.32),* it will be noticed that the ratio S_{yx}/S_y will be large if the coefficient ofdetermination is small, and it will be small when the coefficient of determination is large. Thus

- ✓ if $r^2 = r = 0$, $S_{yx}/S_y = 1$, which means that $S_{yx} = S_y$.
- ✓ if $r^2 = r = 1$, $S_{yx}/S_y = 0$, which means that $S_{yx} = 0$.
- ✓ when $r = 0.865$, $S_{yx} = 0.427\ S_y$ means that S_{yx} is 42.7% of S_y.

Eq. (5.32) also implies that S_{yx} is generally less than S_y. The two can at the most be equal, but only inthe extreme situation when $r = 0$.

Interpretations of r^2:

1. Even though the coefficient of determination, whose under root measures the degree of correlation, is based on S_{yx},; it is expressed as 1 - (S_{yx}/S_y). As it is a dimensionless pure number, the unit in which S_{yx} is measured becomes irrelevant. This facilitates comparison between the two sets of data in terms of their coefficient of determination r^2 (or the coefficient of correlation *r).* This was not possible in terms of $S_{y\ x}$ as the units of measurement could be different.
2. The value of r^2 can range between 0 and 1. When $r^2 = 1$, all the points on the scatter diagram fall onthe regression line and the entire variations are explained by the straight line. On the other hand, when $r^2 = 0$, none of the points on the scatter diagram falls on the regression line, meaning thereby that there is no relationship between the two variables. However, being always non-negative coefficient of determination does not tell us about the direction of the relationship (whether it is positive or negative) between the two variables.
3. When $r^2 = 0.7455$ (or any other value), 74.55% of the total variations in sales are explained by the marketing expenditure used. What remains is the coefficient of non-determination K^2 (= 1 - r^2) = 0.2545. It means 25.45% of the total variations remain unexplained, which are due to factors other than the changes in the marketing expenditure.
4. r^2 provides the necessary link between regression and correlation which are the two related aspects of a single problem of the analysis of relationship between two variables. Unlike regression, correlation quantifies the degrees of relationship between the variables under study, without making a distinction between the dependent and independent ones. Nor does it, therefore, help in predictingthe value of one variable for a

given value of the other.

5. The coefficient of correlation overstates the degree of relationship and it‘s meaning is not asexplicit as that of the coefficient of determination. The coefficient of correlation $r = 0.865$, as compared to $r^2 = 0.7455$, indicates a higher degree of correlation between sales and marketing expenditure. Therefore, the coefficient of' determination is a more objective measure of the degree of relationship.
6. The sum of r and K never adds to one, unless one of the two is zero. That is, $r + K$ can be unity either when there is no correlation or when there is perfect correlation. Except in these two extreme situations, $(r + K) > 1$.

9.2.3 CORRELATION ANALYSIS VERSUS REGRESSION ANALYSIS

Correlation and Regression are the two related aspects of a single problem of the analysis of the relationship between the variables. If we have information on more than one variable, we might be interested in seeing if there is any connection - any association - between them. If we found such a association, we might again be interested in predicting the value of one variable for the given and known values of other variable(s).

1. Correlation literally means the relationship between two or more variables that vary in sympathy so that the movements in one tend to be accompanied by the corresponding movements in the other(s).On the other hand, regression means stepping back or returning to the average value and is a mathematical measure expressing the average relationship between the two variables.
2. Correlation coefficient r_{xy} between two variables X and Y is a measure of the direction and degreeof the linear relationship between two variables that is mutual. It is symmetric, *i.e.*, $r_{yx} = r_{xy}$ and it is immaterial which of X and Y is dependent variable and which is independent variable. Regression analysis aims at establishing the functional relationship between the two(or more) variables under study and then using this relationship to predict or estimate the value of the dependent variable for any given value of the independent variable(s). It also reflects upon the nature of the variable, *i.e.*, which is dependent variable and which is independent variable. Regression coefficient are not

symmetric in X and Y, *i.e.*, $b_{yx} \neq b_{xy}$.

3. Correlation need not imply cause and effect relationship between the variable under study. However, regression analysis clearly indicates the cause and effect relationship between the variables. The variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
4. Correlation coefficient r_{xy} is a relative measure of the linear relationship between X and Y and is independent of the units of measurement. It is a pure number lying between ±1. On the other hand, the regression coefficients, b_{yx} and b_{xy} are absolute measures representing the change in the value of the variable Y (or X), for a unit change in the value of the variable X (or Y). Once the functional form of regression curve is known; by substituting the value of the independent variable we can obtain the value of the dependent variable and this value will be in the units of measurement of the dependent variable.
5. There may be non-sense correlation between two variables that is due to pure chance and has no practical relevance, *e.g.,* the correlation, between the size of shoe and the intelligence of a group of individuals. There is no such thing like non-sense regression.

9.2.4 SOLVED PROBLEMS

Example 9-1

The following table shows the number of motor registrations in a certain territory for a term of 5 years and the sale of motortyres by a firm in that territory for the same period.

Year	Motor Registrations	No. of Tyres Sold
1	600	1,250
2	630	1,100
3	720	1,300
4	750	1,350
5	800	1,500

Find the regression equation to estimate the sale of tyres when the motor registration is known. Estimate sale of tyres when registration is 850.

Solution: Here the dependent variable is number of tyres; dependent on motor registrations. Hence we put motor registrations as X and sales of tyres as Y and we have to establish the regression line of Y on X.

Calculations of values for the regression equation are given below:

X	Y	$d_x = X - \overline{X}$	$d_y = Y - \overline{Y}$	d_x^2	$d_x\, d_y$
600	1,250	-100	-50	10,000	5,000
630	1,100	-70	-200	4,900	14,000
720	1,300	20	0	400	0
750	1,350	50	50	2,500	2,500
800	1,500	100	200	10,000	20,000
$\sum X =$ 3,500	$\sum Y = 6,500$	$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 =$ 27,800	$\sum d_x d_y =$ 41,500

$$\overline{X} = \frac{\sum X}{N} = \frac{3,500}{5} = 700 \quad \text{and} \quad \overline{Y} = \frac{\sum Y}{N} = \frac{6,500}{5} = 1,300$$

b_{yx} = Regression coefficient of Y on X

$$b_{yx} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{\sum d_x d_y}{\sum d_x^2} = \frac{4,1500}{2,7800} = 1.4928$$

Now we can use these values for the regression line

$$Y - \overline{Y} = b_{yx}\ (X - \overline{X})$$

$$\text{or } Y - 1300 = 1.4928\ (X - 700)$$

$$Y = 1.4928\, X + 255.04$$

When $X = 850$, the value of Y can be calculated from the above equation, by putting $X = 850$ in the equation.

$$Y = 1.4928 \times 850 + 255.04$$

$$= 1523.92$$

$$= 1,524 \text{ Tyres}$$

Example 9-2

A panel of Judges A and B graded seven debaters and independently awarded the following marks:

Debator	Marks by A	Marks by B
1	40	32
2	34	39
3	28	26
4	30	30
5	44	38
6	38	34
7	31	28

An eighth debator was awarded 36 marks by judge A, while Judge B was not present. If Judge B were also present, how many marks would you expect him to award to the eighth debator, assuming that the same degree of relationship exists in their judgement?

Solution:

Let us use marks from Judge A as *X* and those from Judge B as *Y*. Now we have to work out the regression line of *Y* on *X* from the calculation below:

Debtor	X	Y	$U = X-35$	$V = Y-30$	U^2	V^2	UV
1	40	32	5	2	25	4	10
2	34	39	-1	9	1	81	-9
3	28	26	-7	-4	49	16	28
4	30	30	-5	0	25	0	0
5	44	38	9	8	81	64	72
6	38	34	3	4	9	16	12
7	31	28	-4	-2	16	4	8
$N = 7$			$\sum U = 0$	$\sum V = 17$	$\sum U^2 = 206$	$\sum V^2 = 185$	$\sum UV = 121$

$$\bar{X} = A + \frac{\sum U}{N} = 35 + \frac{0}{7} = 35 \quad \text{and} \quad \bar{Y} = A + \frac{\sum V}{N} = 30 + \frac{17}{7} = 32.43$$

$$b_{yx} = b_{vu} = \frac{N\sum UV - \left(\sum U \sum V\right)}{N\sum U^2 - \left(\sum U\right)^2} = \frac{7x121\text{-}0x17}{7x206\text{-}0} = 0.587$$

Hence regression equation can be written as

$$Y - \bar{Y} = b_{yx}\,(X - \bar{X})$$

$$Y - 32.43 = 0.587\,(X-35)$$

$$or \quad Y = 0.587X + 11.87$$

When $X = 36$ (awarded by Judge A)

$$Y = 0.587 \times 36 + 11.87$$

$$= 33$$

Thus if Judge B were present, he would have awarded 33 marks to the eighth debater.

Example 9-3 For some bivariate data, the following results were obtained.

Mean value of variable X = 53.2

Mean value of variable Y = 27.9

Regression coefficient of Y on X = - 1.5

Regression coefficient of X on Y = - 0.2

What is the most likely value of Y, when X = 60?

What is the coefficient of correlation between X and Y?

Solution: Given data indicate

$\overline{X}$ = 53.2 $\overline{Y}$ = 27.9

b_{yx} = -1.5 b_{xy} = -0.2

To obtain value of Y for $X = 60$, we establish the regression line of Y on X,

$$Y-\overline{Y} = b_{yx}\ (X-\overline{X})$$

$$Y - 27.9 = -1.5\ (X-53.2)$$

or $$Y = -1.5X + 107.7$$

Putting value of $X = 60$, we obtain

$$Y = -1.5 \times 60 + 107.7$$

$$= 17.7$$

Coefficient of correlation between X and Y is given by G.M. of b_{yx} and b_{xy}

$$r^2 = b_{yx}\, b_{xy}$$

$$= (-1.5) \times (-0.2)$$

$$= 0.3$$

So $$r = \pm\sqrt{0.3} = \pm 0.5477$$

Since both the regression coefficients are negative, we assign negative value to the correlation coefficient

$$r = - 0.5477$$

Example 9-4

Write regression equations of X on Y and of Y on X for the following data

X:	45	48	50	55	65	70	75	72	80	85
Y:	25	30	35	30	40	50	45	55	60	65

Solution: We prepare the table for working out the values for the regression lines.

X	Y	$U = X-65$	$V = Y-45$	U^2	UV	V^2
45	25	-20	-20	400	400	400
48	30	-17	-15	289	255	225
50	35	-15	-10	225	150	100
55	30	-10	-15	100	150	225
65	40	0	-5	0	0	25
70	50	5	5	25	25	25
75	45	10	0	100	0	0
72	55	7	5	49	35	25
80	60	15	15	225	225	225
85	65	20	20	400	400	400
$\sum X = 645$	$\sum Y = 435$	$\sum U = 5$	$\sum V = -20$	$\sum U^2 = 1813$	$\sum V^2 = 1415$	$\sum UV = 1675$

We have, $\bar{X} = \frac{\sum X}{N} = \frac{645}{10} = 64.5$ and $\bar{Y} = \frac{\sum Y}{N} = \frac{435}{10} = 43.5$

$$b_{yx} = \frac{N\sum UV - (\sum U \sum V)}{N\sum U^2 - (\sum U)^2} = \frac{(10)\,x1415 - (5)\,x\,(-20)}{(10)\,x1813 - (5)^2}$$

$$= \frac{14150+100}{18130-25} = \frac{14250}{18105} = 0.787$$

Regression equation of Y on X is

$$Y - \bar{Y} = b_{yx}\,(X - \bar{X})$$
$$Y - 43.5 = 0.787\,(X - 64.5)$$
$$or \quad Y = 0.787X + 7.26$$

Similarly b_{xy} can be calculated as

$$b_{xy} = \frac{N\sum UV - (\sum U \sum V)}{N\sum V^2 - (\sum V)^2} = \frac{(10)\,x1415 - (5)\,x\,(-20)}{(10)\,x1675 - (-20)^2}$$

$$= \frac{14150+100}{16750-400} = \frac{14250}{16350} = 0.87$$

Regression equation of X on Y will be

$$X - \bar{X} = b_{xy}\,(Y - \bar{Y})$$
$$X - 64.5 = 0.87\,(Y - 43.5)$$
$$or \quad X = 0.87Y + 26.65$$

Example 9-5

The lines of regression of a bivariate population are

$$8X - 10Y + 66 = 0$$
$$40X - 18Y = 214$$

The variance of X is 9. Find

(i) The mean value of X and Y

(ii) Correlation coefficient between X and Y

(iii) Standard deviation of Y

Solution: The regression lines given are

$$8X - 10Y + 66 = 0$$
$$40X - 18Y = 214$$

Since both the lines of regression pass through the mean values, the point $(\overline{X}, \overline{Y})$ will satisfy both the equations.

Hence these equations can be written as

$$8\,\overline{X} - 10\,\overline{Y} + 66 = 0$$
$$40\,\overline{X} - 18\,\overline{Y} - 214 = 0$$

Solving these two equations for $\overline{X}$ and $\overline{Y}$, we obtain

$\overline{X} = 13$ and $\overline{Y} = 17$

(ii) For correlation coefficient between X and Y, we have to calculate the values of b_{yx} and b_{xy}

Rewriting the equations

$$10Y = 8X + 66$$
$$b_{yx} = +8/10 = +4/5$$

Similarly, $40X = 18Y + 214$

$$b_{xy} = 18/40 = 9/20$$

By these values, we can now work out the correlation coefficient.

$$r^2 = b_{yx} \cdot b_{xy}$$
$$= 4/5 \times 9/20 = 9/25$$

So $r = \pm\sqrt{9/25}$

$$= \pm 0.6$$

Both the values of the regression coefficients being positive, we have to consider only the positive value of the correlation coefficient. Hence $r = 0.6$

(iii) We have been given variance of X *i.e* $S_x^2 = 9$

$$S_x = \pm 3$$

We consider $S_x = 3$ as SD is always positive

Since $b_{yx} = r\,S_y / S_x$

Substituting the values of b_{yx}, r and S_x we obtain,

$$S_y = 4/5 \times 3/0.6$$
$$= 4$$

Example 9-6: The height of a child increases at a rate given in the table below. Fit the straight line using the method of least-square and calculate the average increase and the standard error of estimate.

Month:	1	2	3	4	5	6	7	8	9	10
Height:	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

Solution:

For Regression calculations, we draw the following table

Month (X)	Height (Y)	X^2	XY
1	52.5	1	52.5
2	58.7	4	117.4
3	65.0	9	195.0
4	70.2	16	280.8
5	75.4	25	377.0
6	81.1	36	486.6
7	87.2	49	610.4
8	95.5	64	764.0
9	102.2	81	919.8
10	108.4	100	1084.0
$\sum X = 55$	$\sum Y = 796.2$	$\sum X^2 = 385$	$\sum XY = 4887.5$

Considering the regression line as $Y = a + bX$, we can obtain the values of a and b from the above values.

$$a = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N\sum X^2 - (\sum X)^2} = \frac{796.2x385 \ -55 \ x4887.5}{10x385-55x55} = 45.73$$

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2} = \frac{10x4887.5-55 \ x796.2}{10x385-55x55} = 6.16$$

Hence the regression line can be written as

$$Y = 45.73 + 6.16X$$

For standard error of estimation, we note the calculated values of the variable against the observed values,

When $X = 1$, $Y_1 = 45.73 + 6.16 = 51.89$

for $X = 2$, $Y_2 = 45.73 + 616 \ x \ 2 = 58.05$

Other values for X = 3 to X = 10 are calculated and are tabulated as follows:

Month (X)	Height (Y)	Y_i	$Y-Y_i$	$(Y-Y_i)^2$
1	52.5	51.89	0.61	0.372
2	58.7	58.05	0.65	0.423
3	65.0	64.21	0.79	0.624
4	70.2	70.37	-0.17	0.029
5	75.4	76.53	-1.13	1.277
6	81.1	82.69	-1.59	2.528
7	87.2	88.85	-1.65	2.723
8	95.5	95.01	0.49	0.240
9	102.2	101.17	1.03	1.061
10	108.4	107.33	1.07	1.145
$\sum(Y_i - Y)^2 = 10.421$				

Standard error of estimation

$$S_{yx} = \sqrt{\frac{1}{N}\sum(Y-Y_i)^2}$$

$$= \sqrt{\frac{10.421}{10}}$$

$$= 1.02$$

Example 9-7: Given $X = 4Y+5$ and $Y = kX + 4$ are the lines of regression of X on Y and of Y on X respectively. If k ispositive, prove that it cannot exceed ¼. If k = 1/16, find the means of the two variables and coefficient of correlation between them.

Solution: Line $X = 4Y + 5$ is regression line of X on Y

So $b_{xy} = 4$

Similarly from regression line of Y on X, $Y = kX + 4$,

We get $b_{yx} = k$

Now

$$r^2 = b_{xy}.\ b_{yx}$$

$$= 4k$$

Since $0 \le r^2 \le 1$, we obtain $0 \le 4k \le 1$,

Or $0 \le k \le \frac{1}{4}$,

Now for $k = \frac{1}{16}$,

$$r^2 = 4x\frac{1}{16} = \frac{1}{4}$$

$$r = +½$$

$$= ½ \text{ since } b_{yx} \text{ and } b_{yx} \text{ are positive}$$

When $k = \frac{1}{16}$, the regression line of Y on X becomes

$$Y = \frac{1}{16}X + 4$$

Or $X - 16Y + 64 = 0$

Since line of regression pass through the mean values of the variables, we obtain revised equations as

$$\overline{X} - 4\overline{Y} - 5 = 0$$

$$\overline{X} - 16\overline{Y} + 64 = 0$$

Solving these two equations, we get

$$\overline{X} = 28 \quad \text{and} \quad \overline{Y} = 5.75$$

Example 9-8 A firm knows from its past experience that its monthly average expenses (X) on advertisement are Rs 25,000 with standard deviation of Rs 25.25. Similarly, its average monthly product sales (Y) have been Rs 45,000 with standard deviation of Rs 50.50. Given this information and also the coefficient of correlation between sales and advertisement expenditure as 0.75, estimate

(i) the most appropriate value of sales against an advertisement expenditure of Rs 50,000.

(ii) the most appropriate advertisement expenditure for achieving a sales target of Rs 80,000

Solution: Given the following

X = Rs 25,000 S_x = Rs 25.25

Y = Rs 45,000 S_y = Rs 50.50

$r = 0.75$

(i) Using equation $Y_c - Y = r\frac{S_y}{S_x}(X - X)$, the most appropriate value of sales Y_c for an advertisement expenditure X = Rs 50,000 is

$$Y_c - 45{,}000 = 0.75\ \frac{50.50}{25.25}(50{,}000 - 25{,}000)$$

$$Y_c = 45{,}000 + 37{,}500$$

$$= \text{Rs } 82{,}500$$

(ii) Using equation $X_c - \overline{X} = r\frac{S_x}{S_y}(Y - \overline{Y})$, the most appropriate value of advertisement expenditure X_c for achieving a sales target Y= Rs 80,000 is

$$X_c - 25{,}000 = 0.75\ \frac{25.25}{50.50}(80{,}000 - 45{,}000)$$

$$X_c = 13{,}125 + 25{,}000$$

$$= \text{Rs } 38{,}125$$

9.3 MULTIPLE REGRESSION ANALYSIS

In multiple regression analysis, the effect of two or more independent variables on one dependent variable is studied. It uses three or more variables to estimate the value of dependent variable. Let's take three variables say X_1, X_2 and X_3. Now, take X_1 as the dependent variable and try to find out its relative movement for movements in both X_2

and X_3 which are independent variables. The prime objectives of multiple regression analysis are:

- To estimate an equation which provides estimates of the dependent variable from the values of the two or more independent variables.
- To obtain a measure of error involved in using regression equation as a basis for estimation.
- To obtain a measure of the proportion of variance in the dependent variable explained by the independent variables

The first objective is accomplished by estimating an appropriate regression equation by the method of least squares. The second objective is achieved through the calculation of a standard error of estimate. The third objective is attained by computing the multiple coefficient of determination.

9.3.1 Regression Equations

A regression equation is an equation for estimating a dependent variable say X_1 from the independent variables X_2, X_3 and is called a regression equation of X_1 on X_2 and X_3. The procedure of estimating multiple regression is similar to the simple regression with the difference that the other variables are added in the regression equation. If there are three variables X_1 X_2 and X_3, the multiple regression has the following form:

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

There are three constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ in the above equation. The subscript after the dot indicatesthe variables which are held constant.

Interpretation of Constants: In the above equation:

$a_{1.23}$ = The constant **$a_{1.23}$** is the intercept made by the regression plane. It gives the value of dependentvariable when X_2 and X_3 independent variables are 0.

$b_{12.3}$ = It indicates the slope of the regression line of X_1 on X_2 when X_3 is held constant. It measures theamount by which a unit change in X_2 is expected to affect X_1 when X_3 is held constant.

$b_{13.2}$ = It indicates the slope of the regression line of X_1 on X_3 when X_2 is held constant. It measures theamount by which a unit change in X_3 is expected to affect X_1 when X_2 is held constant.

In this way, three different regression equations can be formed using three variables X_1, X_2 and X_3which are explained below:

a. The regression plane of X_1 on X_2 and X_3
b. The regression plane of X_2 on X_1 and X_3
c. The regression plane of X_3 on X_1 and X_2

The regression plane of X_1 on X_2 and X_3

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

In the above equation, the value of the value of $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneouslythe following three normal equations:

$$\sum X_1 = Na_{1.23} + b_{12.3}\sum X_2 + b_{13.2}\sum X_3$$
$$\sum X_1X_2 = a_{1.23}\sum X_2 + b_{12.3}\sum X^2_2 + b_{13.2}\sum X_2X_3$$
$$\sum X_1X_3 = a_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X^2_3$$

The last two of these three equations can be obtained if we multiply first equation by X_2 and X_3 on bothsides.

The regression plane of X_2 on X_1 and X_3

$$X_2 = a_{2.13} + {}_{21.3}X_1 + b_{23.1}X_3$$

In the above equation, the value of the value of $b_{21.3}$ and $b_{23.1}$ are determined by solving simultaneouslythe following three normal equations:

$$\sum X_2 = Na_{2.13} + b_{21.3}\sum X_1 + b_{23.1}\sum X_3$$
$$\sum X_1X_2 = a_{2.13}\sum X_1 + b_{21.3}\sum X^2_1 + b_{23.1}\sum X_1X_3$$
$$\sum X_2X_3 = a_{2.13}\sum X_3 + b_{21.3}\sum X_1X_3 + b_{23.1}\sum X^2_3$$

The last two of these three equations can be obtained if we multiply first equation by X_1 and X_3 on bothsides.

The regression plane of X_3 on X_1 and X_2

$$X_3 = {}_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$$

In the above equation, the value of the value of $b_{31.2}$ and $b_{32.1}$ are determined by solving simultaneouslythe following three normal equations:

$$\sum X_3 = Na_{3.12} + b_{31.2}\sum X_1 + b_{32.1}\sum X_2$$

$$\sum X_1X_3 = a_{3.12}\sum X_1 + b_{31.2}\sum X^{2}_{1} + b_{32.1}\sum X_1X_2$$

$$\sum X_2X_3 = a_{3.12}\sum X_2 + b_{31.2}\sum X_1X_2 + b_{32.1}\sum X^{2}_{2}$$

The last two of these three equations can be obtained if we multiply first equation by X_1 and X_2 on both sides.

9.3.2 Assumptions of Linear Multiple Regression Analysis

Followings are the main assumption of linear multiple regression for point estimation:

1. The dependent variable is a random variable whereas the independent variables need not to berandom variables.
2. There must be a linear relationship between the several independent variables and the one dependentvariable.
3. The variance of the conditional distributions of the dependent variable, given various combinationsof values of the independent variables, are all equal.
4. For internal estimation, an additional assumption is that the conditional distributions for thedependent variable follow the normal probability distribution.

9.3.3 Methods of Multiple Regression Analysis

There are three methods of estimating multiple regression equation when deviations are taken from actual means:

First Method

Let $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ be the actual means of the three variable X_1, X_2 and X_3 respectively. The calculation involved in solving these regression equations are considerably reduced if the deviations of the various variables are taken from their respective means. If $x_1 = (X_1 - \bar{X}_1)$, $x_2 = (X_2 - \bar{X}_2)$ and $x_3 = (X_3 - \bar{X}_3)$. The multiple regression equation takes the following shape:

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

We can find the values of $b_{12.3}$ and $b_{13.2}$ by simultaneously solving the following two normal equations:

$$\sum x_1x_2 = b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3$$

$$\sum x_1x_3 = b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2$$

The values of $b_{12.3}$ and $b_{13.2}$ can also be obtained as follows:

$$b_{12.3} = r_{12.3} \times \frac{\sigma_{1.23}}{\sigma_{2.13}}$$

$$b_{13.2} = r_{13.2} \times \frac{\sigma_{1.32}}{\sigma_{3.12}}$$

Second Method

Multiple Regression Equation of X_1 on X_2 and X_3: It can be expressed as follows:

$$X_1 - \bar{X}_1 = \frac{\sigma_1}{\sigma_2} * \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}(X_2 - \bar{X}_2) + \frac{\sigma_1}{\sigma_3} * \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}(X_3 - \bar{X}_3)$$

Multiple Regression Equation of X_2 on X_3 and X_1: It can be expressed as follows:

$$X_2 - \bar{X}_2 = \frac{\sigma_2}{\sigma_3} * \frac{r_{23} - r_{12}r_{13}}{1 - r_{13}^2}(X_3 - \bar{X}_3) + \frac{\sigma_2}{\sigma_1} * \frac{r_{12} - r_{23}r_{13}}{1 - r_{13}^2}(X_1 - \bar{X}_1)$$

Multiple Regression Equation of X_3 on X_1 and X_2: It can be expressed as follows:

$$X_3 - \bar{X}_3 = \frac{\sigma_3}{\sigma_2} * \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}(X_2 - \bar{X}_2) + \frac{\sigma_3}{\sigma_1} * \frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}(X_1 - \bar{X}_1)$$

Third Method

If σ_1, σ_2, σ_3 are not given, we can calculate S_1, S_2 and S_3 from the given data. Then the regression equations are:

Regression equation of X_1 on X_2 and X_3

$$X_1 - \bar{X}_1 = \frac{S_1}{S_2} * \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}(X_2 - \bar{X}_2) + \frac{S_1}{S_3} * \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}(X_3 - \bar{X}_3)$$

Regression equation of X_2 on X_3 and X_1

$$X_2 - \bar{X}_2 = \frac{S_2}{S_3} * \frac{r_{23} - r_{12}r_{13}}{1 - r_{13}^2}(X_3 - \bar{X}_3) + \frac{S_2}{S_1} * \frac{r_{12} - r_{23}r_{13}}{1 - r_{13}^2}(X_1 - \bar{X}_1)$$

Regression equation of X_3 on X_2 and X_1

$$X_3 - \bar{X}_3 = \frac{S_3}{S_2} * \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}(X_2 - \bar{X}_2) + \frac{S_3}{S_1} * \frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}(X_1 - \bar{X}_1)$$

9.3.4 Standard Error of Estimates

The reliability of the estimates obtained through multiple regression equations is studied through the calculations of standard error of the estimate. It is an estimate of unexplained variations in the values of dependent variable X_1. If the total variations of X_1 is divided into two parts, the standard error would represent the unexplained variations. The explained variations would be due to independent variables. The standard error of the estimate of multiple regression equation X_1 on X_2 and X_3 is calculated by the following formula:

$$S_{1.23} = \sqrt{\frac{\sum(X_1 - Y_1)^2}{N-3}}$$

Where $S_{1.23}$ is the standard error of the estimate of X_1 on X_2 and X_3. X_1 is the original value of X and Y_1 is the estimated value on the basis of the regression equations.

The standard error of the estimate in terms of the correlation coefficients r_{12}, r_{13} and r_{23} is:

$$S_{1.23} = \sigma_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$$

If it is proposed not to calculate the estimated values of X for all data points and also not to calculate the partial correlation coefficients, the standard error of the regression estimate is calculate by the following formula:

$$S_{1.23} = \sigma \sqrt{\frac{\sum X_1^2 - a \sum X_1 - b_1 \sum X_2 X_1 - b_2 \sum X_2 X_1}{N-3}}$$

9.3.5 SOLVED EXAMPLES OF MULTIPLE REGRESSION ANALYSIS

Example 9.9 Find the multiple linear regression equation of X_1 on X_2 and X_3 by using normalequations from the data given below:

X_1	2	4	6	8
X_1	3	5	7	9
X_1	4	6	8	10

Solution: **The regression equation of X_1 on X_2 and X_3 is:**

$$X_1 = {}_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \text{ ..... (A)}$$

Where the value of three constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are obtained by solving the following three normal equations:

$$\sum X_1 = Na_{1.23} + b_{12.3}\sum X_2 + b_{13.2}\sum X_3$$

$$\sum X_1X_2 = a_{1.23}\sum X_2 + b_{12.3}\sum X^{2}_{2} + b_{13.2}\sum X_2X_3$$

$$\sum X_1X_3 = a_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X^{2}_{3}$$

Table: Calculation of Required Values

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_1^2	X_2^2	X_3^2
2	3	4	6	8	12	4	9	16
4	5	6	20	24	30	16	25	36
6	7	8	42	48	56	36	49	64
8	9	10	72	80	90	64	81	100
$\sum X_1$ =20	$\sum X_2$ =24	$\sum X_3$ =28	$\sum X_1X_2$ =140	$\sum X_1X_3$ =160	$\sum X_2X_3$ =188	$\sum X_1^2$ =120	$\sum X_2^2$ =164	$\sum X_3^2$ =216

Substituting the values in the normal equations, we get

$$6a_{1.23} + 24b_{12.3} + 28b_{13.2} = 20 \text{..... (i)}$$

$$24a_{1.23} + 164b_{12.3} + 188b_{13.2} = 140 \text{..... (ii)}$$

$$28a_{1.23} + 188b_{12.3} + 216b_{13.2} = 160 \text{ .....(iii)}$$

Multiplying equation (i) by 4 and Subtracting it from the equation (ii), we get:

$$68b_{12.3} + 76b_{13.2} = 60 \text{ ..... (iv)}$$

Multiplying equation (ii) by 7 and equation (iii) by 6, we get:

$$168a_{1.23} + 114b_{12.3} + 1316b_{13.2} = 980 \text{ ……(v)}$$

$$168a_{1.23} + 1128b_{12.3} + 1296b_{13.2} = 960 \text{ ……(vi)}$$

Subtracting (vi) from (v), we obtain:

$$20b_{12.3} + 20b_{13.2} = 20 \text{ ..... (vii)}$$

Multiply equation (iv) by 5 and equation (vii) by 7 to get:

$$340b_{12.3} + 380b_{13.2} = 300 \text{..... (viii)}$$

$$340b_{12.3} + 340b_{13.2} = 340 \text{ .....(ix)}$$

Subtracting (ix) from (viii), we get

$$40b_{13.2} = -40 = b_{13.2} = -1\ldots \text{ ..... (x)}$$

Substitute the value of $b_{13.2}$ in equation (vii), we have:

$$20b_{12.3} + 20\ (-1) = 20$$

$$20b_{12.3} - 20 = 20$$

$$20b_{12.3} = 40$$

$$b_{12.3} = 40/20 = 2 \text{ .....(xi)}$$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (i), we get:

$$6a_{1.23} + 24(2) + 28(-1) = 20$$

$$6a_{1.23} + 48 - 28 = 20$$

$$6a_{1.23} + 20 = 20$$

$$6a_{1.23}$$

$$= 20 –$$

$$20a_{1.23} =$$

$$0/6 = 0$$

Putting the value of $a_{1.23} = 0$, $b_{12.3} = 2$ and $b_{13.2} = -1$, in equation (A) we get the required equation of X_1on X_2 and X_3 as:

$$X_1 = 0$$

$$+ 2X_2 –$$

$$X_3X_1 =$$

$$2X_2 –$$

$$X_3$$

It should be noted that in the above problem X_1, X_2, X_3 are linear and as such the estimated values of $X_1$for given value of X_2 and X_3 would remain unchanged.

Example 9.10

In a trivariate distribution: $\sigma_1 = 3$; $\sigma_2 = 4$; $\sigma_3 = 5$; $r_{23} = 0.4$; $r_{31} = 0.6$; $r_{12} = 0.7$.

Determine the regression of X_1 on X_2 and X_3, when variates are measured from then means.

Solution: The regression equation of X_1 on X_2 and X_3, when variates are measured from then mean is:

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad \text{(i)}$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} * \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} + = \frac{3}{4} \times * \frac{0.7 - (0.6 \times 0.4}{1 - (0.4^2)} + = \frac{0.75 \times 0.46}{0.84} = \frac{0.345}{0.84} = 0.41$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} * \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} + = \frac{3}{5} \times * \frac{0.6 - (0.7 \times 0.4}{1 - (0.4^2)} + = \frac{0.6 \times 0.32}{0.84} = \frac{0.192}{0.84} = 0.229$$

Substituting $b_{12.3} = 0.41$ and $b_{13.2} = 0.229$ in the equation (i), we get the required regression equation of X_1 on X_2 and X_3 as:

$$x_1 = 0.41\, x_2 + 0.229\, x_3$$

Example 9.11

The given data are: $X_1 = 6$, $X_2 = 7$, $X_3 = 8$; $\sigma_1 = 1$, $\sigma_2 = 2$, $\sigma_3 = 3$; $r_{12} = 0.6$, $r_{13} = 0.7$, $r_{23} = 0.8$. Find the regression equation of X_3 on X_1 and X_2. Estimate the value when $X_1 = 4$, and $X_2 = 5$.

Solution: We shall solve this problem by the third method discussed earlier. According to this method, the regression equation of X_3 on X_1 and X_2 is:

$$X_3 - \bar{X}_3 = \frac{\sigma_3}{\sigma_2} * \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}(X_2 - \bar{X}_2) + \frac{\sigma_3}{\sigma_1} * \frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}(X_1 - \bar{X}_1)$$

Substituting the given values in (i) we have:

$$\Rightarrow X_3 - 8 = \frac{3}{2}\left[\frac{0.8 - (0.6 \times 0.7)}{1 - (0.6)^2}\right](X_2 - 7) + \frac{3}{1}\left[\frac{0.7 - (0.8 \times 0.6)}{1 - (0.6)^2}\right](X_1 - 6)$$

$$\Rightarrow X_3 - 8 = 1.5\left[\frac{0.8 - 0.42}{0.64}\right](X_2 - 7) + 3\left[\frac{0.7 - 0.48}{0.64}\right](X_1 - 6)$$

$$\Rightarrow X_3 - 8 = 1.5\left[\frac{0.38}{0.64}\right](X_2 - 7) + 3\left[\frac{0.22}{0.64}\right](X_1 - 6)$$

$$\Rightarrow X_3 - 8 = 1.5 \times 0.59375\,(X_2 - 7) + 3 \times 0.34375\,(X_1 - 6)$$

$$\Rightarrow X_3 - 8 = 0.89\,(X_2 - 7) + 1.03\,(X_1 - 6)$$

$$\Rightarrow X_3 - 8 = 0.89\,X_2 - 6.23 + 1.03\,X_1 - 6.18$$

$$\Rightarrow X_3 = 0.89\,X_2 + 1.03\,X_1 - 6.23 - 6.18 + 8$$

$$\Rightarrow X_3 = 0.89\,X_2 + 1.03\,X_1 - 4.41$$

It is the regression equation of X_3 on X_1 and X_2 for estimating the value of X_3.

Estimation of X_3 by putting $X_1 = 4$ and $X_2 = 5$ in the equation we get:

$$\Rightarrow X_3 = 0.89 \times 5 + 1.03 \times 4 - 4.41$$

$$\Rightarrow X_3 = 4.45 + 4.08 - 4.41 = 4.12.$$

Hence $X_3 = 4.12$, When $X_1 = 4$ and $X_2 = 5$.

Example 9.12 The following table shows the corresponding shows the corresponding values of three variables X_1 X_2 and X_3. Find the least square regression equation of X_3 on X_1 and X_2. Estimate X_3 when$X_1 = 10$ and $X_2 = 6$.

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

Solution: The regression equation of X_3 on X_1 and X_2 can be written as follows:

$$X_3 - \bar{X}_3 = \frac{S_3}{S_2}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\right](X_2 - \bar{X}_2) + \frac{S_3}{S_1}\left[\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right](X_1 - \bar{X}_1) \quad \ldots(i)$$

Table: Computation of $\bar{X}_1, \bar{X}_2, \bar{X}_3, S_1, S_2, S_3, r_{12}, r_{13}, r_{23}$

$(X_1 - \bar{X}_1) = x_1$			$(X_2 - \bar{X}_2) = x_1$			$(X_3 - \bar{X}_3) = x_3$					
X_1	x_1	x_1^2	X_2	x_2	x_2^2	X_2	x_3	x_3^2	$x_1 x_2$	$x_1 x_3$	$x_2 x_3$
3	-5	25	16	+9	81	90	+40	1600	-45	-200	+360
5	-3	9	10	+3	9	72	+22	484	-9	-66	+66
6	-2	4	7	0	0	54	+4	16	0	-8	0
8	0	0	4	-3	9	42	-8	64	0	0	+24
12	+4	16	3	-4	16	30	-20	400	-16	-80	+80
14	+6	36	2	-5	25	12	-38	1444	-30	-228	+190
$\sum X_1 = 48$	$\sum x_1 = 0$	$\sum x_1^2 = 90$	$\sum X_2 = 42$	$\sum x_2 = 0$	$\sum x_2^2 = 140$	$\sum X_3 = 300$	$\sum x_3 = 0$	$\sum x_3 = 4008$	$\sum x_1 x_2 = -100$	$\sum x_1 x_3 = -582$	$\sum x_2 x_3 = 720$

$$\bar{X}_1 = \frac{48}{6} = 8; \qquad \bar{X}_2 = \frac{42}{6} = 7; \qquad \bar{X}_3 = \frac{300}{6} = 50$$

$$S_1 = \sqrt{\frac{\sum(X_1 - \bar{X}_1)^2}{N}} = \sqrt{\frac{90}{6}} = \sqrt{15} = 3.87$$

$$S_2 = \sqrt{\frac{\sum(X_2 - \bar{X}_2)^2}{N}} = \sqrt{\frac{140}{6}} = \sqrt{23.33} = 4.83$$

$$S_3 = \sqrt{\frac{\sum(X_3 - \bar{X}_3)^2}{N}} = \sqrt{\frac{4008}{6}} = \sqrt{668} = 25.85$$

$$r_{12} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \times x_2^2}} = \frac{-100}{\sqrt{90 \times 140}} = \frac{-100}{112.25} = -0.891$$

$$r_{13} = \frac{\sum x_1 x_3}{\sqrt{\sum x_1^2 \times x_3^2}} = \frac{-582}{\sqrt{90 \times 4008}} = \frac{-582}{600.599} = -0.969$$

$$r_{23} = \frac{\sum x_2 x_3}{\sqrt{\sum x_2^2 \times x_3^2}} = \frac{720}{\sqrt{140 \times 4008}} = \frac{720}{749.08} = 0.961$$

Substituting these values in the equation (i), we get:

$$X_3 - 50 = \frac{25.85}{4.83}\left[\frac{0.961 - (-0.969 \times -0.891)}{1 - (-0.9)^2}\right](X_2 - 7) + \frac{25.85}{3.87}\left[\frac{-0.969 - (0.961 \times -0.891)}{1 - (-0.9)^2}\right](X_1 - 8)$$

$$=> \quad X_3 - 50 = 2.546\,(X_2 - 7) - 3.664\,(X_1 - 8)$$

$$=> \quad X_3 - 50 = 2.546\,X_2 - 17.822 - 3.664\,X_1 + 29.312$$

$$=> \quad X_3 = 2.546\,X_2 - 3.664\,X_1 + 61.49$$

It is required regression equation of X_3 on X_1 and X_2.

Estimation of value X_3 putting value of $X_1 = 10$ and $X_2 = 6$ in the equation, we get:

$$=> \quad X_3 = 2.546 \times 6 - 3.664 \times 10 + 61.49$$

$$=> \quad X_3 = 15.276 - 36.64 + 61.49 = 40.126$$

Hence, When $X_1 = 10$, $X_2 = 6$, then $X_3 = 40.126$.

9.4 CHECK YOUR PROGRESS

1. The sum of the squares of the vertical deviations of the observed Y values from the straight line is

   …… .

2. The value of both the regression coefficients cannot be… .....than 1.
3. The mean of both the regression coefficients is either equal to or greater than the …….
4. Regression coefficients are independent of change of .....but not of change of scale.
5. r^2 provides the necessary link between regression and correlation which are the two related aspects ofa single problem of the analysis of between two variables.

9.5 SUMMARY

Regression analysis means the estimation or prediction of the unknown value of one variable from the known value(s) of the other variable(s). It is one of the most important and widely used statistical techniques in almost all sciences - natural, social or physical. Regression analysis for studying more than two variables at a time is known as multiple regressions. If *X* and *Y* are two variables of which relationship is to be indicated, a line that gives best estimate of *Y* for any value of *X*, it is called Regression line of Y on X. If the dependent variable changes to *X*, then best estimate of *X* by any value of *Y* is called Regression line of X on Y. A line of best fit can be

fitted by means of: Free hand drawing method, and least square method. The two regression lines indicate the nature and extent of correlation between the variables. Correlation and Regression are the two related aspects of a single problem of the analysis of the relationship between the variables. If we have information on more than one variable, we might be interested in seeing if there is any connection - any association - between them. If we found such a association, we might again be interested in predicting the value of one variable for the given and known values of other variable(s).

9.6 KEYWORDS

Dependent variable: The variable to be predicted is called the dependent variable.

Independentvariable: The predictor is called the independent variable, or explanatory variable.

The line of Regression: It is the graphical or relationship representation of the best estimate of onevariable for any given value of the other variable.

Error of estimate: The preciseness of an estimate can be obtained only through a measure of themagnitude of error in the estimates, called the error of estimate.

Coefficient of determination: It gives the percentage variation in the dependent variable that isaccounted for by the independent variable.

9.7 SELF-ASSESSMENT TEST

1. Explain clearly the concept of Regression. Explain with suitable examples its role in dealing with business problems.
2. What do you understand by linear regression?
3. What is meant by _regression‘? Why should there be in general, two lines of regression for each bivariate distribution? How the two regression lines are useful in studying correlation between two variables?
4. Why is the regression line known as line of best fit?
5. Write short note on
   (i) Regression Coefficients

(ii) Regression Equations

(iii) Standard Error of Estimate

(iv) Coefficient of Determination

(v) Coefficient of Non-determination

6. Distinguish clearly between correlation and regression as concept used in statistical analysis.
7. Fit a least-square line to the following data:

   (i) Using *X* as independent variable

   (ii) Using *X* as dependent variable

X	:	1	3	4	8	9	11	14
Y	:	1	2	4	5	7	8	9

   Hence obtain

   a) The regression coefficients of *Y* on *X* and of *X* on *Y*

   b) $\bar{X}$ and $\bar{Y}$

   c) Coefficient of correlation between and X and Y

   d) What is the estimated value of Y when X = 10 and of X when Y = 5?

8. What are regression coefficients? Show that $r^2 = b_{yx.} b_{xy}$ where the symbols have their usualmeanings. What can you say about the angle between the regression lines when *(i)* $r = 0$, *(ii)* $r = 1$

   (ii) r increases from 0 to 1?

9. Obtain the equations of the lines of regression of *Y* on *X* from the following data.

X	:	12	18	24	30	36	42	48
Y	:	5.27	5.68	6.25	7.21	8.02	8.71	8.42

   Estimate the most probable value of *Y*, when $X = 40$.

10. The following table gives the ages and blood pressure of 9 women.

Age (*X*) :	56	42	36	47	49	42	60	72	63
Blood Pressure(*Y*)	147	125	118	128	145	140	155	160	149

Find the correlation coefficient between *X* and *Y*.

(i) Determine the least square regression equation of *Y* on *X*.

(ii) Estimate the blood pressure of a woman whose age is 45 years.

11. Given the following results for the height (X) and weight (Y) in appropriate units of 1,000 students:

$\bar{X} = 68, \quad \bar{Y} = 150, \quad S_x = 2.5, \quad S_y = 20$ and $r = 0.6$.

Obtain the equations of the two lines of regression. Estimate the height of a student A who weighs200 units and also estimate the weight of the student B whose height is 60 units.

12. From the following data, find out the probable yield when the rainfall is 29″.

	Rainfall	Yield
Mean	25″	40 units per hectare
Standard Deviation	3″	6 units per hectare

Correlation coefficient between rainfall and production = 0.8.

13. A study of wheat prices at two cities yielded the following data:

	City A	**City B**
Average Price	Rs 2,463	Rs 2,797
Standard Deviation	Rs 0.326	Rs 0.207

Coefficient of correlation r is 0.774. Estimate from the above data the most likely price of wheat

(i) at City A corresponding to the price of Rs 2,334 at City B

(ii) at city B corresponding to the price of Rs 3.052 at City A

14. Find out the regression equation showing the regression of capacity utilization on production fromthe following data:

	Average	**Standard Deviation**
Production (in lakh units)	35.6	10.5
Capacity Utilization (in percentage)	84.8	8.5

$r = 0.62$. Estimate the production, when capacity utilisation is 70%.

15. The following table shows the mean and standard deviation of the prices of two shares in a stockexchange.

Share	**Mean (in Rs)**	**Standard Deviation (in Rs)**
A Ltd.	39.5	10.8
B Ltd.	47.5	16.0

If the coefficient of correlation between the prices of two shares is 0.42, find the most likely priceof share A corresponding to a price of Rs 55, observed in the case of share B.

16. Find out the regression coefficients of Y on X and of X on Y on the basis of following data:

$\sum X = 50, \quad \bar{X} = 5, \quad \sum Y = 60, \quad \bar{Y} = 6, \quad \sum XY = 350$

Variance of $X = 4$, Variance of $Y = 9$

17. Find the regression equation of X and Y and the coefficient of correlation from the following data:

$\sum X = 60, \sum Y = 40, \sum XY = 1150, \sum X^2 = 4160, \sum Y^2 = 1720$ and $N = 10$.

18. By using the following data, find out the two lines of regression and from them compute the KarlPearson's coefficient of correlation.

$\sum X = 250, \sum Y = 300, \sum XY = 7900, \sum X^2 = 6500, \sum Y^2 = 10000, N = 10$

19. The equations of two regression lines between two variables are expressed as

$2X - 3Y = 0$ and $4Y - 5X\text{-}8 = 0$.

(i) Identify which of the two can be called regression line of Y on X and of X ~~on~~ Y.*(ii)*Find X and Y and correlation coefficient r from the equations

20. If the two lines of regression are $4X - 5Y + 30 = 0$ and $20X - 9Y - 107 = 0$ Which of these is the lines of regression of X and Y. Find r_{xy} and S_y when $S_x = 3$

21. The regression equation of profits (X) on sales (Y) of a certain firm is $3Y - 5X + 108 = 0$. The average sales of the firm were Rs 44,000 and the variance of profits is $9/16^{th}$ of the variance of sales. Find the average profits and the coefficient of correlation between the sales and profits.

22. Explain the concept of multiple regression, and try to find out an example in practical field where the multiple regression analysis is likely to be helpful.

23. Explain multiple regression. How does it different from partial regression? Explain with the help ofexample.
24. Given $r_{12} = 0.8$, $r_{13} = 0.7$, $r_{23} = 0.6$, $\sigma_1 = 10$, $\sigma_2 = 8$, $\sigma_3 = 5$. Determine the regression equation of x_2 on x_1 and x_3. (Hint: Here $b_{21.3} = 0.596$, $b_{23.1} = 0.125$)

9.8 ANSWERS TO CHECK YOUR PROGRESS

1. Minimum
2. Greater
3. Coefficient of correlation
4. Origin
5. Relationship

9.9 REFERNCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management *by* G.C. Beri. Tata McGraw Hills Publishing Company Ltd., NewDelhi.
3. Business Statistics *by* Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill PublishingCompany Ltd., New Delhi.
4. Statistics for Business and Economics *by* R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics *by* S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method *by* S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management *by* Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt.Ltd., New Delhi.
8. Statistics for Business and Economics *by* Kohlar Heinz. Harper Collins., New York.

NOTES

NOTES