1

INTRODUCTION TO STATISTICS

Unit Structure

1.0 OBJECTIVES:

After going through this chapter you will able to know:

- Definition of statistics
- Function and Scope of statistics.
- Importance of statistics in real life.
- Limitation of statistics.

1.1 INTRODUCTION

In our daily life, we come across many situation which are directly or indirectly related to numbers.

- Average percentage of students.
- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to smoke.
- In economics, graph of demand and price relation.
- Business growth rates any company.
- One in every 8 South Africans is HIV positive.
- Suppose you need to find the number of employed citizens in a city. If the city has a population of 10 lakh people, we will take a sample of 1000 people. Based on this, we can prepare the data, which is the statistic.
- you may want to predict the price of a stock in six months from now on the basis of company performance measures and other economic factors.

- As a college student, you may be interested in knowing the dependence of the mean starting salary of a college graduate, based on your GPA.

These are just some examples that highlight how statistics are used in our modern society. To figure out the desired information for each example, you need data to analyze.

The purpose of this course is to introduce you to the subject of statistics as a science of data. There is data abound in this information age; how to extract useful knowledge and gain a sound understanding of complex data sets has been more of a challenge. In this course, we will focus on the fundamentals of statistics, which may be broadly described as the techniques to collect, clarify, summarize, organize, analyze, and interpret numerical information.

This course will begin with a brief overview of the discipline of statistics and will then quickly focus on descriptive statistics, introducing graphical methods of describing data. You will learn about combinatorial probability and random distributions, the latter of which serves as the foundation for statistical inference. On the side of inference, we will focus on both estimation and hypothesis testing issues. We will also examine the techniques to study the relationship between two or more variables; this is known as regression.

By the end of this course, you should gain a sound understanding of what statistics represent, how to use statistics to organize and display data, and how to draw valid inferences based on data by using appropriate statistical tools.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life.

Definition:

In simple words statistics is the study and manipulation of given data. It deals with the analysis and computation of given numerical data. Let us take into consideration some more definitions of statistics given by different authors here:

The Merriam-Webster dictionary defines the term statistics as “The particular data or facts and conditions of a people within a state - especially the values that can be expressed in numbers or in any other tabular or classified way”.

According to Sir Arthur Lyon Bowley, statistics is defined as "Numerical statements of facts or values in any department of inquiry placed in specific relation to each other".

Statistics is a branch of mathematics that deals with the collection, review, and analysis of data. It is known for drawing the conclusions of data with the use of quantified models. Statistical analysis is a process of collecting and evaluating data and summarizing it into mathematical form.

Statistics can be defined as the study of the collection, analysis, interpretation, presentation, and organization of data. In simple words, it is a mathematical tool that is used to collect and summarize data.

Uncertainty and fluctuation in different fields and parameters can be determined only through statistical analysis. These uncertainties are determined by the probability that plays a very important role in statistics.

Basics of Statistics

Statistics consist of the measure of central tendency and the measure of dispersion. These central tendencies are actually the mean, median, and mode and dispersions comprise variance and standard deviation.

Mean is defined as the average of all the given data. Median is the central value when the given data is arranged in order. The mode determines the most frequent observations in the given data.

Variation can be defined as the measure of spread out of the collection of data. Standard deviation is defined as the measure of the dispersion of data from the mean and the square of the standard deviation is also equal to the variance.

Mathematical Statistics

Mathematical statistics is the usage of Mathematics to Statistics. The most common application of Mathematical statistics is the collection and analysis of facts about a country: its economy, and, military, population, number of employed citizens, GDP growth, etc. Mathematical techniques like mathematical analysis, linear algebra, stochastic analysis, differential equation, and measure-theoretic probability theory are used for different analytics.

Since probability uses statistics, Mathematical Statistics is an application of Probability theory.

For analyzing the data, two methods are used:

1. Descriptive Statistics: It is used to synopsize (or summarize) the data and their properties.
2. Inferential Statistics: It is used to get a conclusion from the data.

In descriptive statistics, the data or collection of data is described in the form of a summary. And the inferential stats are used to explain the descriptive one. Both of these types are used on a large scale.

There is one more type of statistics, in which descriptive statistics are transitioned into inferential stats.

1.2 FUNCTIONS AND SCOPE:

Function of Statistics:

1. It presents facts in numerical figures:

We can represent the things in their true form with the help of figures. Without a statistical study, our ideas would be vague and indefinite. The facts are to be given in a definite form. If the results are given in numbers, then they are more convincing than if the results are expressed on the basis of quality.

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

2. It presents facts in aggregated and simplified form:

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

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

3. It facilitates comparison:

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

4. Formulation and Testing hypothesis:

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

5. Forecasting:

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

6. Policy making or decision making:

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

7. Its enlarge knowledge:

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

8. To measure uncertainty

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

Scope of Statistics:

Statistics can be used in many major fields such as psychology, geology, sociology, weather forecasting, probability, and much more. The main purpose of statistics is to learn by analysis of data, it focuses on applications, and hence, it is distinctively considered as a mathematical science.

Applications of Statistics

Information around the world can be determined mathematically through Statistics. There are various fields in which statistics are used:

1. **Mathematics:** Statistical methods like dispersion and probability are used to get more exact information.

2. **Business:** Various statistical tools are used to make quick decisions regarding the quality of the product, preferences of the customers, the target of the market etc.

3. **Economics:** Economics is totally dependent on statistics because statistical methods are used to calculate the various aspects like employment, inflation of the country. Exports and imports can be analysed through statistics.

4. **Medical:** Using statistics, the effectiveness of any drug can be analysed. A drug can be prescribed only after analysing it through statistics.

5. **Quality Testing:** Statistics samples are used to test the quality of all the products a Company produces.

6. **Astronomy:** Statistical methods help scientists to measure the size, distance, etc. of the objects in the universe.

7. **Banking:** Banks have several accounts to deposit customers' money. At the same time, Banks have loan accounts as well to lend the money to the customers in order to earn more profit from it. For this purpose, a statistical approach is used to compare deposits and the requesting loans.

8. **Science:** Statistical methods are used in all fields of science.

9. **Weather Forecasting:** Statistical concepts are used to compare the previous weather with the current weather so as to predict the upcoming weather.

There are various other fields in which statistics is used. Statistics have a number of applications in various fields in Mathematics as well as in real life. Some of the major uses of statistics are given below:

- Applied statistics, theoretical statistics, and mathematical statistics
- Machine learning and data mining
- Statistical computing
- Statistics is effectively applied to the mathematics of the arts and sciences
- Used for environmental and geographical studies
- Used in the prediction of weather

1.3 IMPORTANCE:

The student's aims in his study of Statistics:

1. To master the vocabulary of statistics: in order to read and understand a foreign language, there is always the necessity of building up an adequate vocabulary. To the beginner, statistics should be regarded as

a foreign language. The vocabulary consists of concepts that are symbolized by words and by letter symbols.

2. To acquire, or to revive, and to extend skill in computation: Statistics aims at developing computational skills within the students. The understanding of statistics concepts comes largely through applying them in computing operations.

3. To learn to interpret statistical results correctly: Statistical results can be useful only to the extent that they are correctly interpreted. With full and proper interpretation extracted from data, statistical results are the most powerful source of meaning and significance. Inadequately interpreted, they may represent something worse than wasted effort. Erroneously understood they are worse than useless.

4. To grasp the logic of statistics: Statistics provides a way of thinking as well as a vocabulary and a language. It is a logical system, like all mathematics, which is peculiarly adaptable to the handling of scientific problems. Guilford has rightly remarked, " well-planned investigations always include in their design clear considerations of the specific statistical operations to be employed."

5. To learn where to apply statistics and where not to: while all statistical devices can illuminate data, each has its limitations. It is in this respect that the average student will probably suffer most from lack of mathematical background, whether he realizes it or not. Every statistics is developed as a purely mathematical idea. As such, it rests upon certain assumptions. If those assumptions are true of the particular data with which we have to deal, the statistical may be appropriately applied.

6. To understand the underlying mathematics of statistics: This objective will not apply to all students. But it should apply to more than those with unusual previous mathematical training. This will give him a more than commonsense understanding of what goes on in the use of formulas.

1.4 LIMITATIONS

We are known that Statistics is very important tool to study any kind of data but while study statistics we have some limitation as given below.

1. **Qualitative aspect ignored:** The statistical methods don't study the nature of phenomenon which cannot be expressed in quantitative terms. Such phenomena cannot be a part of the study of statistics. These include health, riches, intelligence etc. it needs conversion of qualitative data into quantitative data. So experiments are being undertaken to measure the reactions of a man through data. Now a days statistics is used in all the aspects of the life as well as universal activities.

2. **It does not deal with individual items:** it is clear from the definition given by Prof. Horace Sacrist, " by statistics we mean aggregates of facts and placed in relation to each other", that statistics deals with only aggregates of facts or items and it does not recognize any individual items. Thus, individual terms as death of 6 persons in a accident, 85% results of a class of a school in particular year, will not amount to statistics as they are not placed in a group of similar items. It does not deal with the individual items, however, important they may be.

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

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

5. **Law are not exact:** As far as two fundamental laws are concerned with statistics, i) law of inertia of large number and ii) law of statistical regularity, are not as good as their science laws. They are based on probability. So these results will not always be as good as of scientific laws. On the basis of probability or interpolation, we can only estimate the production of paddy in 2008 but cannot make a claim that it would be exactly 100%. Here only approximations are made.

6. **Results are true only on average:** As discussed above, here the results are interpolated for which time series or regression or probability can be used. These are not absolutely true. If average of two sections of students in statistics is same, it does not mean that all the 50 students is section A has got same marks as in B. There may be much variation between the two. So we get average results. "Statistics largely deals with averages and these averages may be made up of individual items radically different from each other." – W.L. King.

7. **To many method to study problems:** In this subject we use so many methods to find a single result. Variation can be found by quartile deviation, mean deviation or standard deviations and results vary in each case. " it must not be assumed that the statistics is the only

method to use in research, neither should this method of considered the best attack for the problem." – Croxten and Cowden.

8. **Statistical results are not always beyond doubt:** Although we use many laws and formulae in statistics but still the results achieved are not final and conclusive. As they are unable to give complete solution to a problem, the result must be taken and used with much wisdom. "Statistics deals only with measurable aspects of things and therefore, can seldom give the complete solution to problem. They provide a basis for judgment but not the whole judgment."-Prof. L.R Connor.

   Hence statistics is very useful tool but it depends on how the statistician used it more effectively. It depends on the requirement of data analysis.

1.5 LET US SUM UP:

In this chapter we have learn:

- Definition of statistics.
- Functions and Scope of statistics.
- Importance of study of statistics.
- Limitations of statistics.

1.6 UNIT END EXERCISES:

1. What is Statistics? Explain its various uses.
2. Discuss important of Statistics.
3. Give various applications of statistics in Business and Economics.
4. Discuss limitation of Statistics.
5. Explain briefly the functions of Statistics.

Multiple Choice Questions:

1) Statistics is applied in
   a) Economics b) Business management
   c) Commerce and industry d) All these
2) Which of the following is a branch of statistics?
   a) Descriptive statistics b) Inferential statistics
   c) Industry statistics d) Both A and B
3) Which of the following statement is false?
   a) Statistics is derived from the latin word 'Stastu'.
   b) Statistics is derived from the Italian word 'Statista'.
   c) Statistics is derived from the French word 'Statistik'.
   d) None of these

4) Statistics is defined in terms of numerical data in the
   a) Singular sense b) Plural sense
   c) Either (a) or (b) d) both (a) and (b)
5) Statistics concerned with
   a) Qualitative information b) Quantitative information
   c) Either (a) or (b) d) both (a) and (b)
6) An attribute is
   a) A qualitative characteristic b) A quantitative characteristic
   c) A measurable characteristic d) All these

1.7 LIST OF REFERENCES:

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

2

STATISTICAL DATA

Unit Structure

2.0 OBJECTIVES:

After going through this chapter you will able to know:

- Different types of data.
- How to collect data?
- Different method of collection of data.
- Difference between primary and secondary data
- Sample survey and census.

2.1 INTRODUCTION:

The statistical methods or techniques are applicable only when some data are available irrespective of the methods or data collection. The data are collected either by experiments or by survey method and they are tabulated and analysed statistically. Whatever may be the resulting value obtained from analysis, proper and correct interferences have to be drawn from these numerical values. These inferences lead to a final decision.

Our society is highly dependent on data, which underscores the importance of collecting it. Accurate data collection is necessary to make informed business decisions, ensure quality assurance and keep research integrity. During data collection, the statistician must identify the data types, the sources of data and what methods are being used. We will soon see that there are many different data collection methods. There is heavy reliance on data collection in research, commercial, and government fields.

Before an analyst begins collecting data, we can break up data into quantitative and qualitative types. Qualitative data covers descriptions such

as color, size, quality and appearance. Quantitative data unsurprisingly, deals with numbers such as statistics, poll numbers, percentages etc.

Some basic terminology: In statistics as well as in quantitative methodology, the set of data are collected and selected from a statistical population with the help of some defined procedures. There are two different types of data sets namely, population and sample.

Population: In statistics, population is the entire set of items from which you draw data for a statistical study. It can be a group of individuals, a set of items, etc. it makes up the data pool for a study. Generally, population refers to the people who live in a particular area at a specific time. But in statistics, population refers to data on your study of interest. It can be a group of individuals, objects, events, organizations, etc. For e.g. in statistical study we may have a population of a number of students in college.

For the above situation, it is easy to collect data. The population is small and willing to provide data and can be contacted. The data collected will be complete and reliable.

Sample: A sample is defined as a smaller and more manageable representation of a larger group. A subset of a larger population that contains characteristics of that population. A sample is used in statistical testing when the population size is too large for all members or observations to be included in the test.

Variable: A characteristic from the population which can be expressed numerically and which varies from object to object is called a variate. For e.g. weight of students, wages of employees can be measured quantitatively and so these are variates.

Attribute: Certain characteristics cannot be expressed quantitatively but they can be described qualitatively. For e.g. intelligence, beauty, sex, etc. These are called attributes.

Parameter: A statistical measure like mean, standard deviation which is calculated for all objects in the population is called a parameter.

2.2 RELEVANCE OF DATA:

In today's world everything runs on data. Be it from social media to large companies. The term data refers to information about anything. Each company and each institution has a set of data to be maintained and analysed to improve and evaluate the growth of the companies. Analysis of the data or in other words, data analytics is a vast field and one of the most important fields to cover today.

The term data analytics refers to the analysis of the data collected to draw out certain conclusions required as per the company's or researcher objective. It involves the structuring of a massive amount of irregular data and deriving the useful required information from them using statistical

tools. It all involves the preparation of charts, graphs etc. The application of data analytics is not limited to manufacturing companies or any industrial areas, but it get involved in almost every field of human living.

In today's world are moving towards the digital economy, companies have to access to more data than ever before. This data creates a foundation of intelligence for important business decisions. To ensure employees have the right data for decision making, companies must invest in data that improve visibility, reliability, security and scalability.

As application of statistics involves data and therefore the question comes in mind that how to collect data and what are the sources of data?

Data represents information collected in the form of numbers and text. Data collection is generally done after the experiment or observation. Primary data and secondary data are helpful in planning and estimating. Data collection is either qualitative or quantitative.

Different types of data collection methods are used in business and sales organisations to analyse the outcome of a problem, arrive at a solution, and understand a company's performance. Furthermore, there are two types of data collection methods, namely, primary data collection and secondary data collection methods.

There are two categories of data namely, i) primary data and ii) secondary data.

2.3 PRIMARY DATA:

The data which are collected from the units or individual respondents directly for the purpose of certain study or information are known as primary data. For instance, an enquiry is made from each tax payer in a city to obtain their opinion about the tax collecting machinery. The data obtained in a study by the investigator are termed as primary data. If an experiment conducted to know the effect of certain fertilizer doses on the yield or the effect of a drug on the patients, the observations taken on each plot or patient constitute the primary data.

The primary data is the information collected by researcher or investigator for the purpose of the enquiry for the first time. The following are the methods using which the primary data can be collected.

i) Direct personal investigation

ii) Indirect oral investigation

iii) Questionnaires and schedules

Direct personal investigation: In this method the investigator directly meet to person and collect data personally using the following method.

a) Personal contact Method: As the name says, the investigator himself goes to the field, meets the respondents and gets the required information. Here investigator personally interviews the respondent

either directly or through phone or through electronic media. This method is suitable when the scope of investigation is small and greater accuracy is needed.

b) **Telephonic interviewing:** In the present age of communication explosion, telephones and mobile phones are extensively used to collect data from the respondents. This saves the cost and time of collecting the data with a good amount of accuracy.

Indirect Method Investigation: The indirect method is used in cases where it is delicate or difficult to get the information from the respondents due to unwillingness or indifference. The information about the respondent is collected by interviewing the third party who knows the respondent well.

Instances for this type of data collection include information on addiction, marriage proposal, economics status, witnesses in court, criminal proceeding etc. the shortcoming of this method is genuineness and accuracy of the information, as it completely depends on the third party.

Local correspondents: In this method the investigator appoints local agents or correspondents in different places. They collect the information on behalf of the investigator in their locality and transmit the data to the investigator or headquarters. This method is adopted by newspapers, government agencies and trading concerns. This method is less accurate but quick and more expensive.

Questionnaires and Schedules: A questionnaire contains a sequence of questions relevant to the study arranged in a logical order. Preparing a questionnaire is a very interesting and challenging job and required good experience and skill. Questionnaires include open-ended questions and close-ended questions allow the respondent considerable freedom in answering. However, questions are answered in details. Close-ended questions have to be answered by the respondent by choosing an answer from the set of answers given under a question just by ticking.

Before starting the investigation, a question sheet is prepared which is called schedule. The schedule contains all the questions which would extract a complete information from a respondent. The order of questions the language of the questions and the arrangement of parts of the schedule are not changed. However the investigator can explain the questions if the respondent faces any difficulty. It contains direct questions as well as question in tabular form.

Following are the essential of a good questionnaire:

1. The length of questionnaire should be proper one and limited.
2. The language used should be easy and simple. It should not convey two meanings.
3. The term used in questionnaire are explained properly.
4. The questions should be arranged in a proper way.
5. The questions should be in logical manner.

6. The questions should be in analytical form.
7. Complex questions should be broken into filter questions.
8. The questions should be described precisely and correctly.
9. The questionnaire should be constructed for a specific period of time.
10. The questions should be moving around the theme of the investigator.
11. In questionnaire personal questions should be avoided as far as possible.
12. The answers should be short, simple, accurate and direct one .

2.3.1 Editing the primary data:

Once a preliminary draft of the questionnaire has been designed, the researcher is obligated to critically evaluate and edit, if needed. This phase may seem redundant, given all the careful thoughts that went into each question. But recall the crucial role played the questionnaire. The following points must be remembered.

1. The main objective of editing is to detect possible errors and irregularities.
2. While editing primary data the following considerations need attention.
3. The data should be complete.
4. The data should be accurate.
5. The data should consistent.
6. The data should be homogeneous.

Advantages of primary data:

- Data collected is very specific to the problem and its useful.
- Quality of the data collected is not doubtful and is meaningful.
- It may lead to the discovery of additional data and information during its collection.
- It is more accurate and it can be edited or update afterwards.

Disadvantage of primary data:

- There are numerous hassles involved in the collection of primary data like taking a decision such as how, when, what and why to collect.
- The cost involved in the collection of data is very high.
- The collection of primary data is more time consuming.

2.4 SECONDARY:

The data collected through various published or unpublished sources by certain people or agency is known as secondary data. Now information contained in it is used again from records, processed and statistically

analysed to extract some information for other purpose, is termed as secondary data. Such data are cheaper and more quickly obtainable than the primary data and also may be available when primary data cannot be obtained at all.

Types of Sources of Secondary Data:

Secondary data can be obtained from different sources:

1) **Published sources:** Secondary data is usually gathered from published (printed) sources. A few major sources of published information are as follows:

   - Published articles of local bodies, and central and state governments.
   - Statistical synopses, census records, and other reports issued by the different departments of the government.
   - Official statements and publications of the foreign governments
   - Publications and reports of chambers of commerce, financial institutions, trade associations, etc.
   - Magazines, journals, and periodicals.
   - Publications of government organizations like the Central Statistical Organization (CSO), National Sample Survey Organization (NSSO).
   - Reports presented by research scholars, bureaus, economists, etc.

2) **Unpublished sources:** Statistical data can be obtained from several unpublished references.

   - Some of the major unpublished sources from which secondary data can be gathered are as follows:
   - The research works conducted by teachers, professors, and professionals
   - The records that are maintained by private and business enterprises
   - Statistics are maintained by different departments and agencies of the central and the state government, undertakings, corporations, etc.
   - There are various secondary sources of data collection. Some of these include:
   - **Books, Magazines, and Newspapers:** Newspapers, and magazines also carry out surveys and interviews of their own on various aspects like socio-economic conditions, crimes in the country, etc.

- **Reports:** Industries and trade associations also publish reports periodically which contain data regarding trade, production, exports, imports, and the like. The information in these reports will facilitate different types of secondary research.
- **Publications by Renowned Organisations:** Organisations like WHO, ICMR, and other renowned national and international bodies carry out timely surveys and case studies of their own which they then publish on their websites. The data and statistics in these surveys can be accessed by almost everyone by visiting their official website.
- **Research Articles:** Several websites publish research papers by scholars and scientists from respective fields like medicine, finance, economics, etc., which act as secondary data information.
- **Government Data:** Data released by the government of any country is one of the largest sources of secondary data. Sometimes, the central or state government sets up committees to look into some issues. These committees publish reports based on their investigation, which function as a valuable source of secondary data.

- **Advantage of secondary data:**
  - It is less expensive. It saves efforts and time.
  - It helps to make primary data collection more specific since with the help of secondary data, we are able to make out what are the gaps and deficiencies and what additional information needs to be collected.
  - It helps to improve the understanding of the problem.
  - The fact that much information exists in documented form.
  - Many existing data sets are enormous, and far greater than the researcher would be able to collect him or herself, with a far larger sample.
- **Disadvantage of Secondary data:**
  - Accuracy of secondary data is not known.
  - Data may be outdated.
  - Collecting primary data builds up more research skills than collecting secondary data.
  - The researcher has no control over the quality of the data.
  - The data cannot be edited.

2.5 CENSUS AND SAMPLE SURVEY:

- **Census:** A statistical investigation in which the data are collected for each and every element or unit of the population is termed as census method. It is also known as 'complete enumeration' or '100% enumeration' or 'complete survey'. A census method is that process of the statistical list where all members of a population are analysed. The population relates to the set of all observations under concern. For instance, if you want to carry out a study to find out student's feedback about the amenities of your school, then all the students of your school would form a component of the 'population' for your study. In our country, the Government conducts the **Census of India** every ten years. The Census appropriates information from households regarding their incomes, the earning members, the total number of children, members of the family, etc. This method must take into account all the units. It cannot leave out anyone in collecting data.

  But instead of study entire population we can study part of the population is called sample.

- **Sample survey:** A survey involving the collection of data about sample of units selected from the population is called sample survey. For example if we want to find the mean weight of all BMS students studying in University of Mumbai, we may select a sample of BMS students from IDOL. Here the survey in which the data regarding weights of IDOL BMS students are obtained is a sample survey.

  A sample survey required information about only a fraction of the population and therefore it saves money and time in comparison to census survey. In sample survey the amount of work is reduced and so we can afford and obtain more accurate results.

Steps in involving in a sample survey:

- Each objective should be defined clearly, and relevant questions related to the objective should be introduced in the questionnaire.
- These objectives should be measurable, specific and should be able to help to derive the results that are expected from the survey.
- Your objectives thoroughly and clearly, the next step is to determine the population.
- Selecting a sample from the population is a significant step.
- The next step would be designing the survey. Many of the sample surveys collect different types of information.
- The next step would be to implement the survey to the required sample population and collect the data from them.
- The next step is to analyze the data. Statistically, correct data should be analyzed so that the results are precise.

- The analysis is done by keeping the objectives of the study in mind so that the results that are obtained relevant to the study.
- Data analysis should be done to arrive at proper conclusions that are relevant to the study.

2.6 LET US SUM UP:

In this chapter we have learn:

- Different types of data i.e. primary data and secondary data.
- Different Methods of collection of primary data.
- How to edit primary data.
- Different sources of secondary data.
- Census method of data collection for population.
- Sample survey method for data collection sample.

2.7 UNIT END EXERCISES:

1. Write the short on types of data.
2. Distinguish between primary data and secondary data.
3. Distinguish between the census method and sampling method.
4. What are different methods to collect primary data by direct investigation method?
5. Explain how primary data is edited.
6. Write advantage and disadvantage of primary data.
7. Write advantage and disadvantage of secondary data.
8. Explain the sources of secondary data.
9. Describe the various steps involving to conduct sample survey.
10. What are the requirements of a good questionnaire?
11. What are the various ways of collecting primary data?
12. Write a short note on Census.

Multiple Choice Questions:

1) Data collected on religion from the census reports are
   a) Primary data b) Secondary data
   c) Sample data d) (a) or (b)
2) The primary data collected by
   a) Interview method b) observation method
   c) Questionnaire method d) All these
3) The quickest method to collect primary data is

a) Personal interview b) Indirect interview

c) Telephone interview d) By observation

4) Some important sources of secondary data are

a) International and government sources

b) International and primary sources

c) Private and primary sources

d) Government sources

5) The best method to collect data, In case of a natural calamity, is

a) Personal interview b) Indirect interview

c) Questionnaire method d) Indirect observation metho

2.8 LIST OF REFERENCES:

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

3

PRESENTATION OF DATA

Unit Structure

3.0 OBJECTIVES:

After going through this chapter you will able to know:

- The classification of data to represent the data.
- To represent data we can used tabulation.
- Types of tabulation of data.
- To represent the data we used graphical methods.

3.1 INTRODUCTION:

After collecting, the desired data the first step is to be taken is to classify and tabulate the data. In order to make the data simple and easily understandable, simplify them in such a way that irrelevant data are removed and their significant features are standing out prominently. The procedure adopted for this purpose is known as method of classification and tabulation. The classification and tabulation provide a clear picture of the collected data and on that basis the further processing is decided.

After study of the importance and techniques of classification and tabulation that help to arrange the mass of collected data in a logical and summarize manner. However, it is a difficult and cumbersome task for common man and researcher to interpret the data. Too many figures are often confusing and may fail to convey the message effectively to those for whom it is meant.

To overcome this inconvenience, the most appealing way in which statistical results may be presented is through diagrams and graphs.

A diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationships. If we draw diagrams on the basis of the data collected they will be understood and appreciated by all. Every day we

can find the presentation of stock market, cricket score etc. in newspaper, television and magazines in the form of diagrams and graphs.

In this chapter we will discuss classification, tabulation and some of the major types of diagrams, graphs and maps frequently used in presenting statistical data.

3.2 CLASSIFICATION OF DATA:

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

Usually the data can be collected through questionnaire, schedules or response sheets. This collected data need to be consolidated for the purpose of analysis and interpretation. This process is known as Classification and Tabulation. We can include a huge volume of data in a simple statistical table and one can get an outline about the model by observing the statistical table rather the raw data. To construct diagrams and graphs, it is essential to tabulate the data.

For Example, letters in the post office are classified according to their destinations viz., Delhi, Madurai, Bangalore, Mumbai etc.,

Requisites of Ideal Classification

It should be unambiguous: There should be no uncertainty or ambiguity. Classes should be defined rigidly, so as to avoid any ambiguity.

It should be flexible: The classification should be enough to accommodate change, amendment and inclusion in various classes in accordance with new situations.

It should be homogeneous: Units of each class should be homogeneous. All the units included ia a class or group should be present according to the property on basis of which the classification was done.

It should be suitable for the purpose: The composition of the class should be according to the purpose.

For example: to find out the economic condition of the persons, create classes on the basis of income.

It should be stable: Stability is necessary to make data comparable and to make out meaningful comparison of the results. This means that the classification of data set into different classes must be performed is a way, that whenever an investigation is carried out, there is no change in classes and so the results of the investigation can be compared easily.

It should be exhaustive: Each and every item of data must belong to a particular class. An ideal classification is one that is free from any residual classes such as others or miscellaneous, as they do not state the characteristics clearly and completely.

It should be mutually exclusive: The classes should be mutually exclusive.

Formation of a Discrete Frequency Distribution:

The formation of discrete frequency distribution is quite simple. The number of times a particular value is repeated is noted down and mentioned against that values instead of writing that value repeatedly. In order to facilitate counting prepare a column of tallies. In another column, place all possible values of variable from lowest to the highest. Then put a bar (vertical line) opposite the particular value to which it relates. To facilitate counting, blocks of five bars are prepared and some space is left in between each block. Finally count the number of bars and get the frequency.

Example 1: The daily wages in Rs. Paid to the works are given below. Form the Discrete Frequency Distribution.

300, 200, 200, 100, 100, 200, 100, 100, 400, 100, 100, 200, 400, 300, 100, 300, 100, 100, 200, 100, 100, 100, 200, 100, 200, 100, 100, 100, 400, 200.

Solution: Frequency distribution of daily wages in Rs

Daily wages (in Rs.)	Tally Marks	No. of workers
100	~~IIII IIII IIII I~~	16
200	~~IIII~~ III	08
300	III	03
400	III	03
		Total = 30

Formation of a Continuous Frequency Distribution :

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

a. **Class Limits:** Class limits are the lowest and the highest values that can be included in a class. The two boundaries of class are known as the lower limit and the upper limit of the class. The lower limit of the a class is the value below which there can be no item in the class. The upper limit of the a class is the value above which there can be no item in the class.

For example, for the class 20 – 40, 20 is the lower limit and 40 is the upper limit. If there was an observation 40.5, it would not be included in this class. Again if there was an observation of 19.5, it would not be included in this class.

b. **Class Intervals:** The difference between upper and lower limit of the class is known as class interval of that class.

For example, in the class 20 – 40, the class interval is 20 (i. e. 40 minus 20). An important decision while constructing a frequency

distribution is about the width of the class interval i. e. whether it should be 10, 20, 50, 100, 500 etc. It depends upon the range in the data, i. e. the difference between the smallest and largest item, the details required and number of classes to be formed, etc.

Following is the simple formula to obtain the estimate of appropriate class interval,

$i = \frac{L-S}{k}$,

where i = Class interval,

L = largest item

S = smallest item

K = the number of classes.

For example, if the salary of 100 employees in a company undertaking varied between Rs. 1000 to 6000 and we want to form 10 classes then the class interval would be

$i = \frac{L-S}{k}$

L = 6000, S = 1000, k = 10

$i = \frac{L-S}{k} = \frac{6000-1000}{10} = 500$

The staring class would be 500 – 1000, 1000 – 1500 and so on.

The question now is how to fix the number of classes i.e. k. The number can be either fixed arbitrarily keeping in view the nature of problem under study or it can be decided with the help of Sturge's Rule.

According to Sturge's Rule, number of classes can be determined by the formula:

k = 1 + 3.322 log N

where N = total number of observations and

log = logarithms of the number.

Therefore, if 10 observations are there, the number of classes shall be,

k = 1 + (3.322 x 1) = 4.322 or 4 [∵ log 10 = 1]

Therefore, if 100 observations are there, the number of classes shall be,

k = 1 + (3.322 x 2) = 7.644 or 8 [∵ log 100 = 2]

It should be noted that since log is used in the formula, the number of classes shall be between 4 and 20. It cannot be less than 4 even if N is less than 10 and if N is 10 lakh, k will be 1 + (3.322 x 6) = 20.9 or 21.

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

d. Class Mid-point or Class mark: Mid-point of a class-interval is calculated for further calculations in statistical work.

$$\text{Mid-point of a class} = \frac{Upper\ limit\ of\ the\ class + Lower\ limit\ of\ the\ class}{2}$$

Methods of Data Classification:

There are two methods of classifying the data according to class intervals.

a. Exclusive method: When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class it is known as the exclusive method of classification.

Weight (in kg)	No. of Students
40-50	40
50-60	160
60-70	110
70-80	200
80-90	90

In the above example, there are 40 students whose weight is between 40 to 49.99 kg. A student whose weight is 50 kg is included in the class 50 – 60.

b. Inclusive method: In this method, the upper limit of one class is included in that class itself.

Weight (in kg)	No. of Students
40-49	40
50-59	160
60-69	110
70-79	200
80-89	90

In the class 40 – 49, we include students whose weight is between 40 kg and 49 kg. If the weight of the student is exactly 50 kg he is included in next class.

Example 2: Prepare a frequency distribution for the students marks data.

25	85	41	70	85	55	85	55	72
72	50	90	52	68	72	52	91	53
79	75	60	35	65	80	70	70	36
66	55	80	72	41	88	60	45	78
42	90	66	47	80	88	91	82	50
52	55	72	68	65				

Solution: Since the lowest value is 25 and the largest value is 91, we take class intervals of 10.

Marks	Tally Marks	Frequency
25 – 35	I	01
35 – 45	~~IIII~~	05
45 – 55	~~IIII~~ III	08
55 – 65	~~IIII~~ I	06
65 – 75	~~IIII IIII~~ IIII	14
75 – 85	~~IIII~~ II	17
85 – 95	~~IIII~~ IIII	19
		Total = 50

Example 3: Prepare a frequency distribution for the following data by taking class interval such that their mid values are 17, 22, 27, 32 and so on.

30	30	36		33	42	27	22	41	30	42
30	21	54		36	31	40	28	19	48	26
48	15	37		16	17	54	42	51	44	32
42	31	21		25	36	22	41	40	46	52

Solution: Since we have to classify the data in a such manner that the mid values are 17, 22, 27, 32 and so on The first class interval should be 15 – 19 (mid-value = (15 + 19)/2 = 17).

Variable	Tally Marks	Frequency
15 – 19	IIII	4
20 – 24	IIII	4
25 – 29	IIII	4
30 – 34	~~IIII~~ III	8
35 – 39	IIII	4
40 – 44	~~IIII~~ IIII	9
45 – 49	III	3
50 – 54	IIII	4
		Total = 40

3.3 TABULATION:

The simplest and most revealing devices for summarizing data and presenting them in a meaningful manner is the statistical table. After classifying the statistical data, next step is to present them in the form of tables. A table is a systematic organization of statistical data in rows and columns. The purpose of a table is to simplify the presentation and to facilitate comparisons. The main objective of tabulation is to answer various queries concerning the investigation. Tables are very helpful for doing analysis and drawing inferences from them.

Classification and tabulation go together, classification being the first step in tabulation. Before the data are put in tabular form, they have to be classified.

Objectives of Tabulation

- **To simplify complex data:** It reduces raw data in a simplified and meaningful form. The reader gets a very clear idea of what the table present. It can be easily interpreted by a common person in less time.
- **To facilitate comparison:** Since the table is divided into rows and columns, for each row and column there is total and subtotal, the relationship between different parts of data can be done easily.
- **To bring out essential features of data**: It brings out main features of data. It presents facts clearly and precisely without textual explanation.
- **To give identity to the data**: when the data are arranged in a table with title and number, they can be differently identified.
- **To save space**: Table saves space without sacrificing the quality and quantity of data.

Parts of Table

Generally, a table should be comprised of the following components:

1. **Table Number:** Each table must be given a number. Table number helps in distinguishing one table from other tables. Usually tables are numbered according to the order of their appearance in a chapter. For example, the first table in the first chapter of a book should be given number 1.1 and second table of the same chapter be given 1.2. Table number should be given at its top or towards the left of the table.

2. **Title of the Table:** The title is a description of the contents of the table. Every table must be given suitable title. A complete title has to answer the questions what categories of statistical data are shown, where the data occurred and when the data occurred. The title should be clear and brief. It is placed either just below the table number or at its right.

3. **Caption:** Caption refers to the column headings. It may consist of one or more column headings. Under one column there may be sub heads. The caption should be clearly defined and placed at the middle of the column. If the different columns have different units, the units should be mentioned with the captions.

4. **Stub:** Stub refers to the rows or row heading. They are at extreme left of the table. The stubs are usually wide than column headings but they are as narrow as possible.

5. **Body:** It is most important part of the table. It contains number of cells. Cells are formed by intersection of rows and columns. The body of the table contains numerical information.

6. **Headnote:** It is used to explain certain points relating to the whole table that have not included in the title, in the caption or stubs. It is placed below the title or at the right hand corner of the table. For example, the unit of measurement is frequently written as a headnote, such as "in thousands", "in crores", etc.

7. **Footnotes:** It helps in clarifying the point which is not clear from the title, captions or stubs. It is placed at the bottom of a table.

There are different ways of identifying the footnotes. One is numbering them consecutively with small numbers 1, 2, 3 or letters a, b, c, d. Another way identifies the first footnote with one star (*), second footnote with two stars (**), third footnote with three stars (***) and so on. Sometimes instead of * , +,@,£ etc used.

3.4 GRAPHICAL PRESENTATION OF DATA:

Graphical presentations are very simple for even a common person to understand. It is popular method of presentation of data. With the help of graphs, two or more sets of data can be easily compared and analysed. The trend of the data also can be seen from the graph.

A graph is drawn in a plane with two reference lines called the X-axis (horizontal) and the Y-axis (vertical). The axes are perpendicular to each other and their point of intersection is called *Origin*. Every point in the plane is identified by two coordinates (x, y). The first coordinate (x) represents the value of the variable on the X-axis and the second coordinate (y) represents the value of the variable on the Y-axis.

Proper scale of measurement should be taken to accommodate the complete data on the graph. If needed the origin can be shifted from (0, 0) to any other required value. Such a process is called *shifting of origin*.

Now, we shall study different types of graphical presentation of data:

One Dimensional or Bar Diagrams

This is the most common type of diagrams. They are called one-dimensional diagrams because only length of the bar matters and not the width. For large number of observations lines may be drawn instead of bars to save space.

Types of Bar Diagrams:

a. Simple bar diagram

b. Subdivided bar diagram

c. Multiple bar diagram

d. Percentage bar diagram

Simple Bar Diagram: A simple bar diagram is used to represent only one variable. It should be kept in mind that, only length is taken into account and not width. Width should be uniform for all bars and the gap between each bar is normally identical. For example the figures of production. Sales, profits etc for various years can be shown by bar diagrams.

Example 1: Prepare a simple bar diagram for following data related to wheat exports.

Year	Exports (in million tons)
2003	12
2004	15
2005	19
2006	25
2007	40

Solution: By taking years on x axis, Exports (in million tons) on y axis, rectangles of equal width are drawn. The distance between successive rectangles is same. The scale on y axis is 1 cm = 5 million tons.

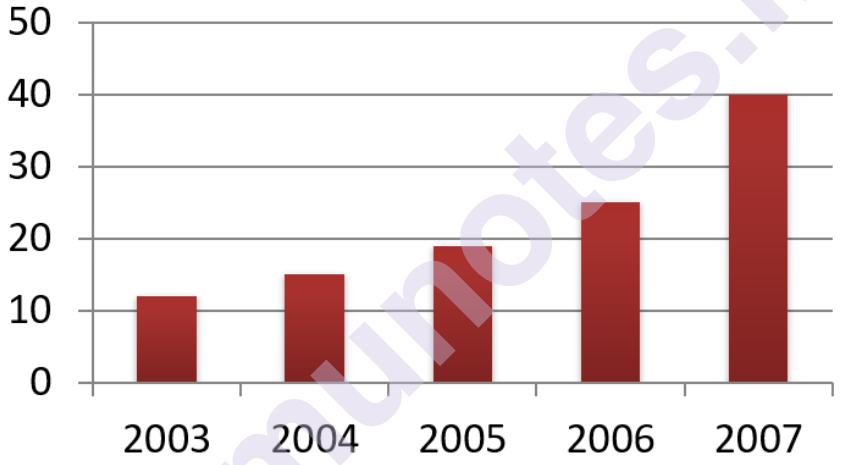


Figure 3.1 Simple Bar Diagram Showing the Wheat Exports in Different Years

Subdivided Bar Diagram: In this diagram, one bar is constructed for total value of the different components of the same variable. Further, it is subdivide impropriation to the various components of that variable.

A bar is represented in the order of magnitude from the largest component at the base of the bar to the smallest at the end of the bar, but the order of various components in each bar is kept in the same order. Different shades or colors are used to distinguish between different components. To explain such differences, the index should be used in the bar diagram. The subdivided bar diagrams can be constructed both on horizontal and vertical bases.

Example 2: The following data shows the production of rice for the period 2010 to 2018. Represent the data by a subdivided bar diagram.

Year	Non-Basmati Rice (in Million metric tons)	Basmati Rice (in Million metric tons)	Total (in Million metric tons)
2010	29	35	64
2011	35	33	68
2012	25	35	60
2013	40	30	70
2014	42	32	74
2015	32	40	72

Solution:

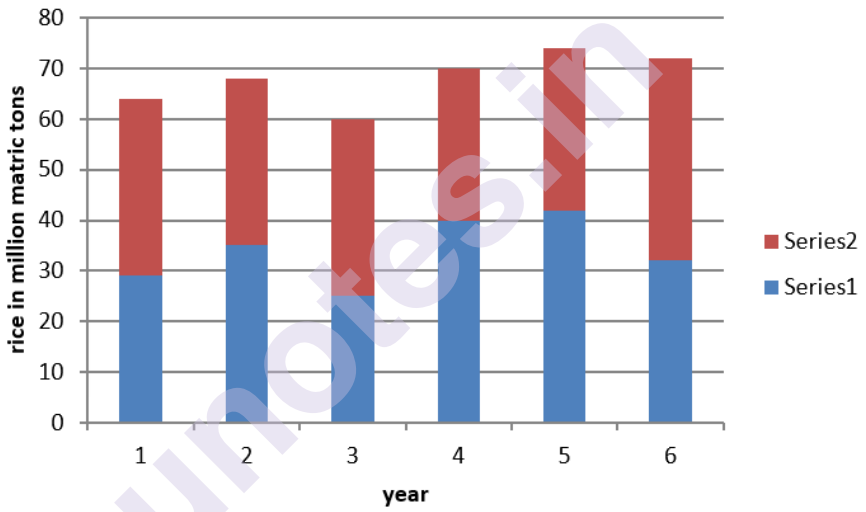


Figure 3.2 Subdivided Bar Diagram Showing the production of Rice (in Different Years)

Multiple Bar Diagram: Whenever the comparison between two or more related variables is to be made, multiple bar diagram should be preferred. In multiple bar diagrams two or more groups of interrelated data are presented. The technique of drawing such type of diagrams is the same as that of simple bar diagram. The only difference is that since more than one components are represented in each group, so different shades, colors, dots or crossing are used to distinguish between the bars of the same group.

Example 3: Represent the following data by a multiple bar diagram.

Class	Physics	Chemistry	Mathematics
Student A	50	63	57
Student B	55	60	68
Student C	48	60	55

Solution:

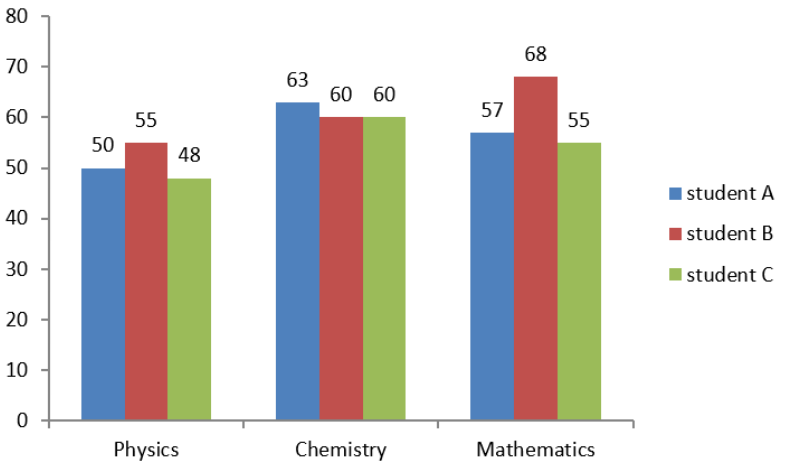


Figure 3.3 Multiple Bar Diagram

Percentage Bar Diagram: Percentage bars are particularly useful in statistical work which requires the representation of the relative changes in data. When such diagrams are prepared, the length of the bars is kept equal to 100 and segments are cut in these bars to represent the percentages of an average.

Example 4: Draw percentage bar diagram for following data.

Particulars	Cost Per Unit (2010)	Cost Per Unit (2020)
Material	22	35
Lobour	30	40
Delivery	10	20
Total	62	95

Solution: Express the values in terms of percentage for both the years.

Particulars	Cost Per Unit (2010)	% Cost	Cumulative % cost	Cost Per Unit (2020)	% Cost	Cumulative % cost
Material	22	35.48	35.48	35	36.84	36.84
Lobour	30	48.38	83.86	40	42.10	78.94
Delivery	10	16.12	99.98	20	21.05	99.99
Total	62	100		95	100	

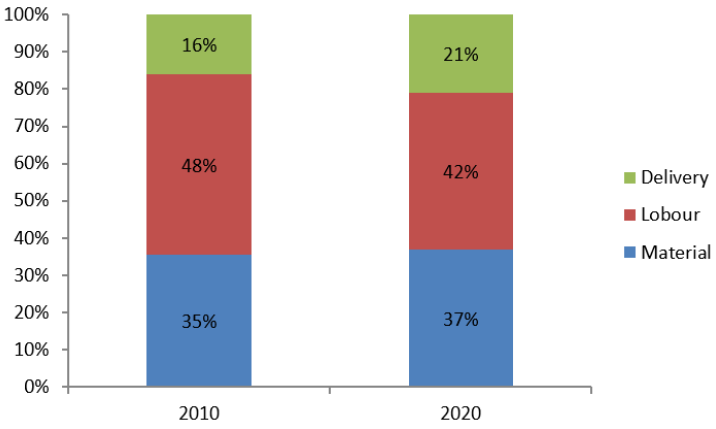


Figure 3.4 Percentage Bar Diagram

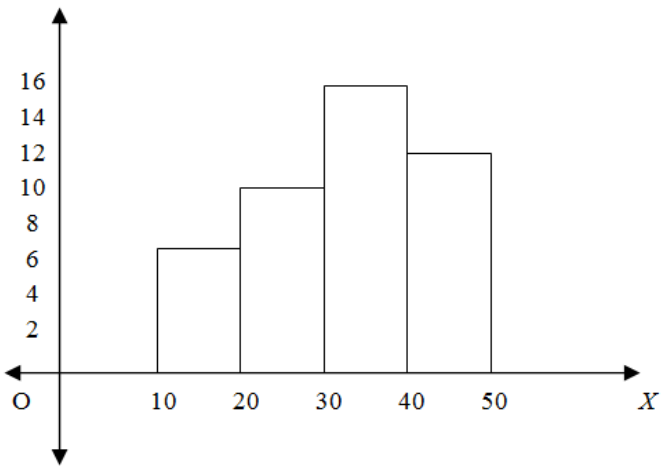
Histogram

A Histogram is a graph of a frequency distribution where adjacent rectangles are drawn to represent the data. The width of the rectangles depends upon the class width of the class intervals which are taken on the X-axis. The height of the rectangles depends upon the class frequency, which are taken on the Y-axis.

Example: Draw a histogram representing the following frequency distribution:

Sales in '000 Rs.	10-20	20-30	30-40	40-50
No of companies	06	10	16	12

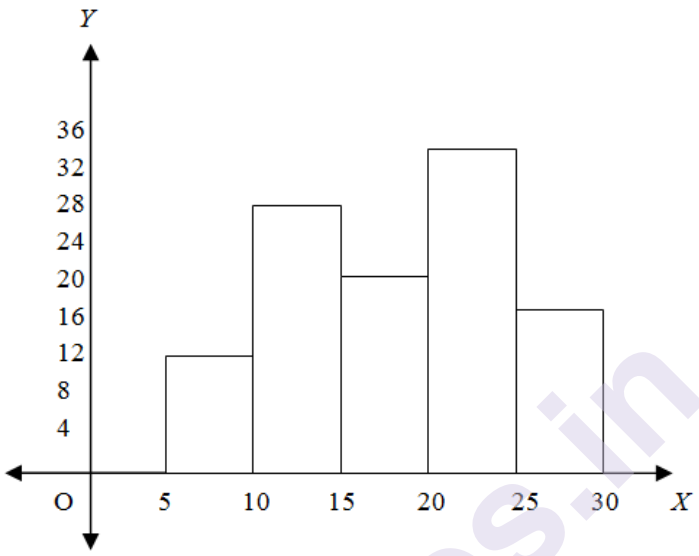
Solution: The class intervals are taken on the X-axis and the frequencies (no of companies) are taken on the Y-axis. The Histogram of the given data is as shown below:



Example 2:Draw a Histogram to present the following data:

Ages in yrs	5 – 10	10 – 15	15 – 20	20 - 25	25 – 30
No of Boys	12	28	20	32	16

Solution:



Note: If the class intervals are discrete they should be first converted to continuous class intervals.

3.4.2 Frequency Curve

To draw a frequency curve, the class marks of the continuous class intervals are computed and taken on the X-axis. The frequencies are taken on the Y-axis. The class marks are plotted against the corresponding frequencies. These points are then joined by a smooth curve. This resultant curve is called as *frequency curve*.

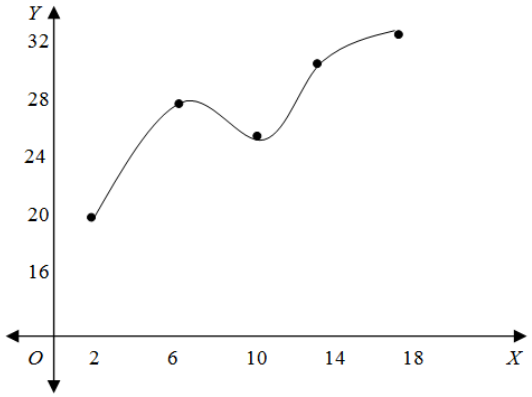
The only important care that has to be taken is that in the process of joining the successive points by smooth curve the trend of the data is not hampered.

Example 3:

Draw a frequency curve for the following data:

Income .	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20
in '000 Rs.					
No of families	20	28	26	30	32

Ans: The class marks of the class intervals are 2, 6, 10, 14 and 18. Now these class marks are plotted against the corresponding class frequencies and the frequency curve is drawn as shown below:



3.4.3 Ogive Curves

Ogive curves are the frequency curves in which instead of class marks the class limits (either upper or lower) are plotted against the cumulative frequencies (either less than or more than type). Hence *Ogive curves* are also called as *cumulative frequency curve*.

Less than Ogive Curve

The *upper class limits* are plotted against the *less than* cumulative frequencies and joined by a smooth curve. A less than Ogive curve is always in upward direction.

More than Ogive Curve

The *lower class limits* are plotted against the *more than* cumulative frequencies and joined by a smooth curve. A more than Ogive curve is always in downward direction.

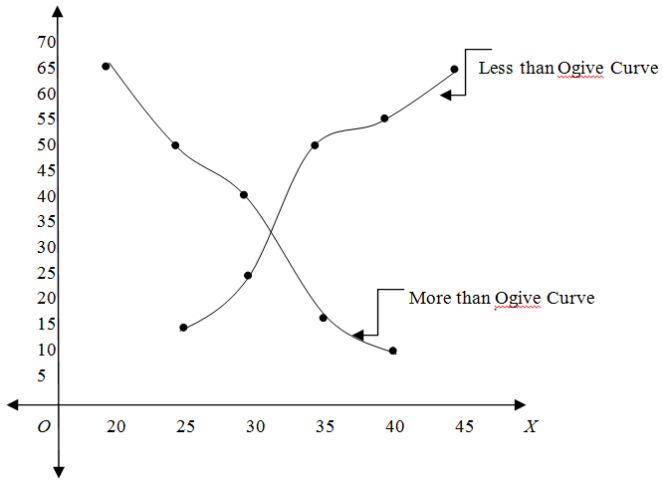
Example 4:

Draw the Ogive Curves for the following data:

Weight in Kg	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
No of children	15	10	25	5	10

Ans: We prepare the less than and more than cumulative frequency table and draw both the Ogive curves as shown below:

Weight in kg	**No of children**	**less than *cf***	**more than *cf***
20 – 25	15	15	65
25 – 30	10	25	50
30 – 35	25	50	40
35 – 40	5	55	15
40 – 45	10	65	10



Ogive Curves are very useful to find graphically the positional averages like median, quartiles, percentiles and mode. We will study all this in the next chapter.

3.5 LET US SUM UP:

In this chapter we have learn:

- Classification of data.
- Types of data classification.
- Tabulation of different types of data.
- Different types of diagrams and graphs to represent the data.

3.6 UNIT END EXERCISES:

1. Define the following terms :

   a) Frequency b) Class – Interval

   c) Class – Limits d) Class – marks

   e) Cumulative Frequency

2. The following data gives the height of 24 students in cm of a class. Prepare a frequency distribution table and find the cumulative frequencies.

   145 146 138 152 144 155 172 160 168 173
   170 140 150 145 165 135 141 153 167 156
   166 174 133 170

3. The following data gives the weight of 30 boundaries in tons. Prepare a frequency distribution table and find (a) relative frequencies and (b) percentage frequencies

12	7	8		15	16	14	11	9	5	13
18	12	6		10	5	4	12	13	17	14
9	14	16		11	12	19	20	5	10	4

4. The following data gives the ages of 50 child labours. Prepare a frequency distribution table with cumulative frequencies and answer the following question (i) how many labours are there whose age is less than 9, (ii) how many child labours have age less than 13, (iii) how many child labours are there with age more than 13? Take class intervals as 5 – 7, 7 – 9, 9 – 11, …….

12	6	5		13	8	15	9	11	14	6
10	14	9		12	11	7	8	13	11	13
9	6	5		10	14	12	5	8	6	13
11	11	10		6	14	13	11	7	9	12
8	11	9		13	12	6	11	8		

5. The rainfall in mm is given to a certain area. Prepare the frequency distribution table. Write down the class marks and class width of the class intervals. Prepare % frequencies and cumulative frequencies for the data.

24.5	16.5	13.8		15.5	19.7	21.3	30	28.4	14.2
17.9	16.5	11.8		13.2	15.4	24.5	26.1	18.6	19.2
27.6	26.8	21.5		24	16.2	17.1	14.5	19.5	23.4
25.8	27.6	18.2							

6. Convert the following inclusive class intervals into exclusive type

   (i) 10 – 14, 15 – 19, 20 – 24, 25 – 29

   (ii) 22 – 28, 30 – 36, 38 – 44, 46 – 52, 52 – 58

   (iii) 2 – 12, 16 – 26, 30 – 40, 44 – 54, 58 – 68

7. Prepare a bivariate frequency distribution table for the following data:

Marks in Statistics	15	10	18	28	20	30	35	45	14	16	40	4	48	28	12
	25	15	35	18	6	26	21	32	45	9	11	16	35	41	34
Marks in Bus. Law	5	20	8	15	6	22	28	39	26	18	38	12	29	38	13
	26	27	9	25	17	30	41	40	18	15	40	5	35	7	12

8. Prepare a bivariate frequency distribution table for the following data:

Height in cm	132	145	130	150	136	147	154	155	132	136	141	146	154	150	152
	135	140	137	151	142	139	148	146	154	145	152	136	151	144	151
Weight in Kg	26	32	38	45	40	35	48	52	36	32	50	40	38	54	58
	27	33	46	57	55	59	60	40	45	32	49	32	51	60	33

9. Write a short note on different types of graphs.
10. Explain briefly the Ogive curves.
11. Distinguish between Histogram and Frequency curve.
12. Draw a histogram for the following data:

C.I.	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30
f	4	10	18	14	20	18

13. Draw a histogram for the following data:

Rainfall in mm	30 – 45	45 – 60	60 – 75	75 – 90	90 – 105	105 – 120
No of Cities	11	6	14	23	16	10

14. Draw a frequency curve for the following data:

Time in min	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12
No of customers	7	10	17	15	16	25

15. Draw Ogive curves for the following data:

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 - 90	90-100
No of students	9	22	33	40	36	54	42	23	17	13

Find from the graph:

1. Number of students who have got distinction (marks more than 75)
2. Number of students who have got marks less than 35. (Hint: For more than 75: On the X-axis locate 75. Draw a perpendicular till it touches more than ogive curve from there extend it on Y-axis. The point where it meet is required answer.)

Multiple Choice Questions:

1) The given data 5,1,0,4,2,4,3,1,2,6 is called :-

a) Frequency distribution data b) Grouped frequency data

c) Row data d) None of the above

2. _____ is used to present data involving one variable.

a) Multiple Bar diagram b) Pie diagram

c) Simple bar diagram d) None of these.

3. The median of a given frequency distribution is found graphically with the help of _____.

a) Histogram b) Simple bar diagram

c) Frequency polygon d) Ogive.

4. The best method of presentation of data is

a) Tabular b) Textual c) Diagrammatic d) (b) and (c)

5. The most accurate mode of data presentation is

a) Diagrammatic methods b) Tabulation

c) Textual presentation d) None of these

6. The frequency distribution of a continuous variable is known as

a) Grouped frequency distribution

b) Simple frequency distribution

c) ungroup frequency distribution

d) None of these

7. Mutually inclusive classification is usually meant for

a) A discrete variable b) A continuous variable

c) An attribute d) All these

8. A comparison among the class frequencies is possible only in

a) Frequency polygon b) Histogram

c) Ogives d) (a) and (b)

9. The number of types of cumulative frequency is

a) one b) Two

c) Three d) Four

3.7 LIST OF REFERENCES:

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

4

MEASURES OF CENTRAL TENDENCY

Unit Structure

4.0 OBJECTIVES:

After going through this chapter you will able to know:

- The concept of central tendency (Average) of data.
- The different measure of central tendency.
- The advantage and disadvantage of mean.
- Compute measure of central tendency for ungroup and group data.
- Calculation of combined mean and weighted mean.

4.1 INTRODUCTION

In statistical analysis there is a need of condensation of the huge data available so as to study its different characteristics. In the previous chapter we have seen how to classify and present the data in a tabular form. The data arranged in the frequency distribution tables shows a tendency to cluster around at certain values. This tendency is called as *Central Tendency* and is calculated statistically. A measure of central tendency or an average is the single value representing the complete data. It is an important summary measure in statistics. The word *average* in Statistics has a quantitative meaning and not qualitative.

As defined by Clark and Sekkade : " An average is an attempt to find out one single figure to describe the whole of figures ".

4.2 OBJECTIVES OF AN AVERAGE

(i) A single value representation of the entire data:

With the help of an average one can present a huge amount of data in a summarized form, which is easier to understand. It gives a bird's eye view of the entire data. For Example, it is not possible to and required to know the individual requirements of petrol consumption, but an average quantity of petrol consumption is enough for the government agencies in planning petroleum imports.

(ii) To compare different statistical data:

Also it helps to compare different sets of data. For Example, the passing percentage of students of two colleges can be compared by the average passing percentage of students of each of the college.

(iii) To analyse and facilitate decision making:

The most important aspect of an average is that it can be used for analyzing the data and hence making some predictions or decisions based on that. For Example, if the average monthly revenue of a product is found to be decreasing, then the manufacturer can think of some advertising or other measures to increase the revenue.

4.3 FUNDAMENTALS OF A GOOD AVERAGE

(*i*) It should be easily understood: Since statistical measures are used for simplifying huge data, an average should be easily understood by the end user.

(*ii*) It should be rigidly defined: As statistical measures are used by different people, an average should be simple but properly or rigidly defined so as to avoid alterations caused due to interpretations of different individuals. There should be no chance of a bias of an individual in its calculation.

(*iii*) It should be easy to calculate: In order to make an average most popular it is important that the algebraic formula and the method is not difficult or complex for an individual to calculate.

(*iv*) It should be based on entire data: An average should be based on each and every observation of a given data. If any observation is deleted it should affect the average also, otherwise it cannot be called as a representative of the entire data.

(*v*) It should not be excessively affected by extreme values: Since an average is a measure of central tendencies for a large data, it should truly represent characteristics of the entire data. Hence, it should not be excessively affected or distorted by the extreme (very small or very large) observations.

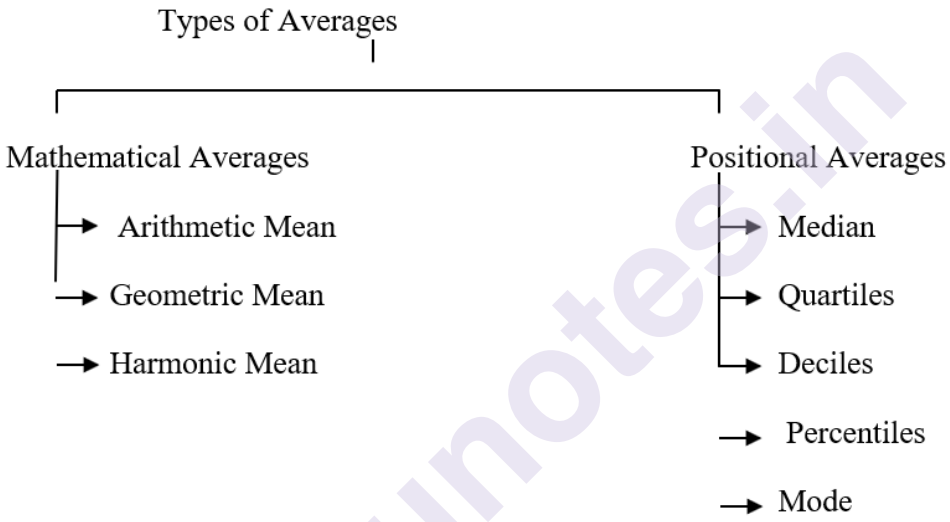
(*vi*) It should be capable of further algebraic treatment: Any statistical measure should be useful for further analyses or calculations. Otherwise it would be of limited use to statisticians.

(*vii*) It should have sampling stability: If independent sets of samples of same size and type of data are taken, we should expect approximately same average for each sample. In other words, an average should have sampling stability.

Now we shall study the different types of averages.

TYPES OF AVERAGES

The various measures of central tendencies can be classified into two major types: (*i*) Mathematical Averages and (*ii*) Positional Averages. These are further classified into subtypes as follows:



As per our scope of syllabus, we will restrict ourselves to arithmetic mean, geometric mean, median, quartiles and mode.

4.4 ARITHMETIC MEAN:

SIMPLE (OR UNWEIGHTED) ARITHMETIC MEAN:

The simple arithmetic mean (A.M.) is defined as the ratio of sum of all the observations to the total number of observations. In symbolic form, let $x_1, x_2, x_3, ..., x_n$ be 'n' number of observations. The A.M. is denoted by $\bar{x}$ and is given by the formula: $\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ or simply $\bar{x} = \frac{\sum x}{n}$.

In order to simplify the formula notation, the index 'i' can be skipped from the summation symbol.

Example 1: The marks of 10 students are as follows: 02, 07, 04, 05, 06, 05, 09, 03, 07, and 04. Find the average marks.

Solution: Average marks $\bar{x} = \frac{2+7+4+5+6+5+9+3+7+4}{10} = \frac{52}{10} = 5.2$

Example 2: The monthly sales (in '00 Rs.) of a product are given below. Find its average monthly sales.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sales	10	14	13	09	08	11	15	16	15	17	16	14

Solution: $\bar{x} = \frac{10+14+13+9+8+11+15+16+15+17+16+14}{12} = \frac{158}{12} = 13.17$

Thus, the average sales are Rs. 1,317.

Example 3: The average score of Rohit in 5 matches is 56, of which 4 matches scores are as 45, 31, 68, and 52. How much his score in fifth match.

Solution: Here, total number of matches, $n = 5$.

Let he score x_5 runs in fifth match.

$$\sum x = 45 + 31 + 68 + 52 + x_5 = 196 + x_5$$

The average score is 56.i.e. $\bar{x} = 56$.

Now

$$\bar{x} = \frac{\sum x}{n} = \frac{196 + x_5}{5}$$

$$56 = \frac{196 + x_5}{5}$$

$$56 \times 5 = 196 + x_5$$

$$x_5 = 280 - 196 = 84.$$

Therefore, he scores 84 runs in fifth match.

Example 3: Ajit has given first year BMS exam and score average marks 65. He attempted 7 papers of which 5 papers marks are as 58, 62, 74, 78, 82. But difference between remaining two subject marks is 7. Find the remaining two subjects marks.

Solution: Here total number of subjects, $n = 7$.

Let remaining two subjects marks are $x_6,\ x_7$.

$$\sum x = 58 + 62 + 74 + 78 + 82 + x_6 + x_7 = 354 + x_6 + x_7$$

Now

$$\bar{x} = \frac{\sum x}{n} = \frac{354 + x_6 + x_7}{7}$$

$$65 = \frac{354 + x_6 + x_7}{7}$$

$$65 \times 7 = 354 + x_6 + x_7$$

$$x_6 + x_7 = 455 - 354 = 101$$

$\therefore x_6 + x_7 = 101$ ……(I)

Given that the difference between two subjects is 7.

i.e. $x_6 - x_7 = 7$…….(II)

Solving equations (I) and (II) we get

$x_6 = 54$ and $x_7 = 47$

Therefore he score 54 and 47 marks in remaining two subjects.

For group data: The group data is also called frequency distribution. We have learn that there are two types of group data i) discrete data and ii) continuous data.

For discrete data: Let the variable X takes values $x_1, x_2, x_3, \ldots\ldots\ldots, x_n$ with the correspondence frequencies $f_1, f_2, f_3, \ldots\ldots\ldots, f_n$ respectively, than the mean of the variable X is given by,

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots\ldots\ldots\ldots + f_nx_n}{f_1 + f_2 + f_3 +, \ldots\ldots\ldots + f_n}$$

$$\bar{x} = \frac{\sum fx}{N} \qquad \because \sum f = N$$

Steps to calculate A.M.:

1. We calculate the total frequency denoted $N = \sum f$, where $\sum f = f_1 + f_2 + f_3 + ... + f_n$
2. Then we calculate the products $f_1.x_1, f_2.x_2, f_3.x_3, \ldots., f_n.x_n$.
3. The A.M. is now calculated by the formula: $\bar{x} = \frac{\sum fx}{N}$, where

$\sum fx = f_1x_1 + f_2x_2 + f_3x_3 + ... + f_nx_n$.

Example 5: The following is the frequency distribution of heights of students in a class in a college. Calculate the average height of the class.

Height (in cms)	152	153	154	155	156	157	158
No. of	10	16	20	28	18	14	14

Students							

Solution: (Students are directed to draw vertical tables while practicing the problems)

Let the height (in cms) be denoted by ***x*** and the no. of students by ***f***.

Introducing the column of ***fx*** we have,

Height (in cms) (***x***)	152	153	154	155	156	157	158	Total
No. of Students (***f***)	10	16	20	28	18	14	14	N = 120
Fx	1520	2448	3080	4340	2808	2198	2212	Σfx =18606

From the table we have: Σfx =18606 and N = 120.

$\therefore \bar{x} = \frac{\Sigma fx}{N} = \frac{18606}{120} = 155.05$. Thus, the average height of the class is 155.05 cms.

Example 6: The following table gives the survey report of 100 people who were asked to count maximum number in one breath. Find the average numbers one can count in one breath.

Numbers	50	60	70	80	90	100	110	120
No. of Persons	25	20	15	5	15	10	5	5

Solution: Let the numbers be denoted by ***x*** and the no. of persons by ***f***. Introducing the column of ***fx***:

Numbers (***x***)	50	60	70	80	90	100	110	120	Total
No. of Persons (***f***)	25	20	15	5	15	10	5	5	N = 100
Fx	1250	1200	1050	400	1350	1000	550	600	$\sum fx$ = 6400

From the table we have: N = 100 and $\sum fx$ = 6400.

The average numbers in one breath is $\bar{x} = \frac{\Sigma fx}{N} = \frac{6400}{100} = 64$.

Grouped (continuous) A.M.

Consider a grouped and continuous class distribution with class intervals as $a_1 - a_2, a_2 - a_3, a_3 - a_4, ...etc$. Let the corresponding frequencies be denoted by $f_1, f_2, f_3, ..., f_n$.

Steps to calculate A.M. for continuous class distributions

1. We calculate the total frequency denoted $N = \sum f$
2. Now the midpoints (x_i) of the continuous class intervals are calculated. For Example, the midpoint of first class interval is $x_1 = \frac{a_1 + a_2}{2}$, for the second interval it is $x_2 = \frac{a_2 + a_3}{2}$ and so on.
3. Now we introduce the column of ***fx***. The midpoints (x_i) are multiplied by the corresponding frequencies (f_i). The sum of these products is calculated and is denoted by $\sum fx$.
4. The A.M. is now calculated by the formula: $\bar{x} = \frac{\sum fx}{N}$.

Example 7: The minimum marks for passing in a post graduate diploma course are 50. The following data gives the marks of 100 students who passed the course. Find the average marks scored by the students.

Marks	50-59	60-69	70-79	80-89	90-99
No. of Students	28	30	25	15	2

Solution: In calculation of mean it is not necessary to convert the inclusive class intervals to exclusive ones. Now, we introduce two columns for calculating the mean. (*i*) mid points of intervals and (*ii*) ***fx***

Marks	Mid points (***x***)	Frequency(***f***)	***fx***
50-59	54.5	28	1526
60-69	64.5	30	1935
70-79	74.5	25	1862.5
80-89	84.5	15	1267.5
90-99	94.5	2	189
Total	--	N = 100	$\sum fx = 6780$

From the table: $\sum fx = 6780$ and N = 100

$$\therefore \bar{x} = \frac{\Sigma fx}{N} = \frac{6780}{100} = 67.8$$

Example 8: The following table gives the marks of 100 students. Find the average marks.

Marks (less than)	10	20	30	40	50	60	70	80	90	100
No. of students	4	12	23	45	60	70	75	82	92	100

Solution: Observe that the data given has marks less than, which means the frequency given is less than cumulative type. Converting them to

frequencies and introducing two columns as done in the previous problem, the mean is calculated as follows:

Marks	Mid points (x)	Frequency(f)	fx
0-10	5	4	20
10-20	15	8	120
20-30	25	11	225
30-40	35	22	770
40-50	45	15	675
50-60	55	10	550
60-70	65	5	325
70-80	75	7	525
80-90	85	10	850
90-100	95	8	760
Total	--	N = 100	$\sum fx = 4820$

The average marks are: $\bar{x} = \frac{\Sigma fx}{N} = \frac{4820}{100} = 48.2$

MISSING FREQUENCIES

Example 9: If the arithmetic mean for the following data is 24.8, find the missing frequency.

Class Interval 0-10 10-20 20-30 30-40 40-50

Frequency 9 11 -- 12 8

Ans: Let the missing frequency be k. Now, completing the table for calculating the mean, we have:

Class Interval	Mid points (x)	Frequency(f)	fx
0-10	5	9	45
10-20	15	11	165
20-30	25	K	$25k$
30-40	35	12	420
40-50	45	8	360
Total	--	N = 40 + k	$\sum fx = 990 + 25k$

Now, given $\bar{x} = 24.8$ and $\bar{x} = \frac{\Sigma fx}{N}$

$\therefore 24.8 = \frac{990 + 25k}{40 + k}$ $\therefore 24.8(40 + k) = 990 + 25k$

$\Rightarrow 992 + 24.8k = 990 + 25k$ $\Rightarrow 0.2k = 2$. Thus, $k = 10$.

The missing frequency is 10.

CORRECTED MEAN

Example 10: The average weight of 25 boys from age group 10-15 was calculated as 38 Kg. Later on it was found that one of the boy's weight was wrongly taken as 34 Kg instead of 43 Kg. Find the corrected mean.

Solution: Given $n = 25$ and $\bar{x} = 38$, wrong value = 34, correct value = 43

The formula for mean is $\bar{x} = \frac{\Sigma x}{n}$, which means $\Sigma x = n.\bar{x}$.

Using this we have $\Sigma x = 25 \text{ x } 38 = 950$. This sum is wrong as one of the observation is incorrect. So we find the correct sum by subtracting the wrong value from the sum and adding the correct value.

Correct sum = Σx – wrong value + correct value = 950 – 34 + 43 = 959

$\therefore$ correct mean = (correct sum)/n = 959/25 = 38.36

4.5 WEIGHTED A.M.

In calculating simple arithmetic mean we assume that all observations are of equal importance. Practically it may be that some observations are more important or less important as compared with the rest. For example, while computing average salary of a company, the class I, class II, class III and class IV employees have different levels of salaries and allowances and hence are not of same level. A simple arithmetic mean in this case will not be representative of all the employees of the company. In such cases weights are assigned to different observations depending upon their importance.

Let $x_1, x_2, x_3, ..., x_n$ be 'n' number of observations with corresponding weights as w_1, w_2, …, w_n.

Then the A.M. is calculated by the formula: $\bar{x} = \frac{w_1x_1 + w_2x_2 + ... + w_nx_n}{w_1 + w_2 + ... + w_n} = \frac{\sum wx}{\sum w}$

Example 11: Compute the weighted mean salary of employees in a company from the following data:

Employee	Class I	Class II	Class III	Class IV
Salary (in '000 Rs.)	30	24	18	10
Number of employees	10	25	10	5

Solution:

Employee	Class I	Class II	Class III	Class IV	
Salary (***x***)	30	24	18	10	Total

(in '000 Rs.)					
Number of employees (*f*)	10	25	10	15	N = 60
Fx	300	600	180	50	$\Sigma fx =$ 1130

$\bar{x} = \frac{\Sigma fx}{N} = \frac{1130}{60} = 18.83$. The weighted average salary is Rs. 18,830.

Example 12: Given below is the performance of students of three colleges A, B and C in different courses. The data gives the percentage of students passed and no. of students (in'000). Using weighted average mean find out the best performing college.

Colleges	College A		College B		College C	
B.A.	65%	2	80%	4	70%	3
B.Com.	54%	3	75%	4	50%	2
B.Sc.	72%	1	70%	2	80%	5

Solution:

Colleges	College A			College B			College C		
Courses	*x*	*w*	*Wx*	*x*	*w*	*wx*	*x*	*w*	*wx*
B.A.	65	2	130	80	4	320	70	3	210
B.Com.	54	3	162	75	4	300	50	2	100
B.Sc.	72	1	72	70	2	140	80	5	400
Total	-	$\sum w = 6$	$\sum wx = 364$	-	$\sum w = 10$	$\sum wx = 760$	-	$\sum w = 10$	$\sum wx = 710$

Now the weighted mean for different college is calculated as follows:

College A

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

$= \frac{364}{6}$

$\therefore \bar{x}_w = 60.67$

College B

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

$= \frac{760}{10}$

$\therefore \bar{x}_w = 76$

College C

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

$= \frac{710}{10}$

$\therefore \bar{x}_w = 71$

Comparing all the three weighted means, since the weighted mean of College B is highest, its performance is the best among the three colleges.

Example 13: In an entrance examination, different weights were attached for the subject Maths, Physics, Chemistry and Biology. The marks obtained by Tejas, Dhyey and Anant are given below. Find the weighted mean and give a comment on it.

Subject	Students			Weight
	Tejas	Dhyey	Anant	
Maths	72	68	81	4
Physics	88	70	79	3
Chemistry	75	74	84	3
Biology	83	72	89	1

Solution: The weighted means for the three students is calculated as follows:

Subject	Weight (w)	Students					
		Tejas		Dhyey		Anant	
		X	wx	x	Wx	x	wx
Maths	4	75	300	68	272	81	324
Physics	3	90	270	70	210	79	237
Chemistry	3	77	231	74	222	84	252
Biology	1	92	92	72	72	89	89
Total	Σw =11	--	Σwx =893	--	Σwx =776	--	Σwx =902

Now we calculate the weighted mean for the three students using the formula: $\bar{x}_w = \frac{\Sigma wx}{\Sigma w}$

For Tejas: $\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{893}{11} = 81.18$

For Dhyey: $\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{776}{11} = 70.55$

For Anant: $\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{902}{11} = 82$.

Comparing the individual weighted mean, since Anant has the highest among the three students, his performance is the best.

In the above case if instead of weighted mean, simple mean is calculated then the conclusion may differ and would not be proper as all the subjects are not to be treated equally.

For Tejas: $\Sigma x = 334 \Rightarrow \bar{x} = \frac{\Sigma x}{n} = \frac{334}{4} = 83.5$

For Dhyey: $\Sigma x = 284 \Rightarrow \bar{x} = \frac{\Sigma x}{n} = \frac{284}{4} = 71$

For Anant: $\Sigma x = 333 \Rightarrow \bar{x} = \frac{\Sigma x}{n} = \frac{333}{4} = 83.25$

Comparing the three means the obvious conclusion is that Tejas's performance is better than of Anant.

4.6 COMBINED A.M.

Any statistical measure is expected to be useful in further algebraic treatment of the data. For Example, if the average daily wages of men and women are known then the average wages for the total workers can be useful for the employer. This can be done as follows: Let $\bar{x}_1$ be the A.M. for a set of n_1 observations and $\bar{x}_2$ be the mean for a set of n_2 observations. Then the combined mean for the set of $n_1 + n_2$ observations is given by the formula : $\bar{x}_c = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$.

This formula can be extended to any 'k' number of sets of observations. The combined mean hence will be given by: $\bar{x}_c = \frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n_1 + n_2 + ... + n_k}$.

Example 14: The mean daily wages of 20 women workers are Rs. 100 and that of 35 men workers are Rs. 140. Find the mean daily wages for all workers taken together.

Solution: The data can be tabulated as follows:

	Men	Women
Number	35	20
Mean daily wages	140	100

Let $n_1 = 35$, $\bar{x}_1 = 140$ and $n_2 = 20$, $\bar{x}_2 = 100$

The combined mean daily wages are:

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{35 \text{ x } 140 + 20 \text{ x } 100}{35 + 20} = \frac{6900}{55} = 125.45$$

Example 15: The average height of students in a class is 162 cm. If the average height of boys is 167 cm and that of girls is 160 cm, find the ratio of boys to girls. Also find the number of boys and girls if there are 80 boys in the class.

Solution: Given, $\bar{x}_c = 162$, $\bar{x}_1 = 167$ and $\bar{x}_2 = 160$. Let there be *a* boys and *b* girls in the class.

Now, $\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$. $\quad \therefore 162 = \frac{167a + 160b}{a + b}$

$\therefore 162a + 162b = 167a + 160b \quad \Rightarrow 2b = 5a$.

$\therefore a : b = 2 : 5$. Thus, the ratio of boys to girls is 2:5

If there are 80 boys in the class, the number of girls = 5 x (80/2) = 200.

4.7 MERITS AND DEMERITS OF ARITHMETIC MEAN

Merits:

1. It is simple to calculate and easy to understand.
2. It is rigidly defined.
3. It can be used for further algebraic treatment very easily and hence is a popular statistical measure.
4. It is based on all observations of a given set of data.
5. It is least affected by the sampling fluctuations.
6. It is useful in comparing two or more sets of data.

Demerits or Limitations

1. If the number of observations is not large, then some extreme items in the data may considerably affect the average. For Example, in a cricket match if one player has scored 200 runs and rest have scored together 100 runs, then the total runs for the team is 300 with an average score for each player being 27 runs. This does not represent the data in true sense as the remaining 10 players actually have an average score of only 10 runs.
2. The A.M. has an upward bias. The large observations dominate the small observations in calculating its average. For example, the average of 1, 1.1, 0.9 and 21 is 6. We can see clearly that the because of the last observation the average has gone up.
3. Comparisons based merely on A.M. may be misleading. For Example, two contractors payment to their workers is suppose 800, 250, 50, 50, 50 and 350, 300, 300, 150, 100. The average pay for both comes out

to be 240, which may force to conclude that both the contractors pay equal wages to their workers, which as we can see is far from true.

4. The A.M. obtained may not be among the observations taken into consideration. For Example, the A.M. of 10, 20, and 70 is 33.33, which is not among the three observations. A.M. being a mathematical average, it does not reflect the qualitative part like sincerity, beauty etc of the data. Also, if there are 100, 110 and 125 students in three classes then the average number of students is 111.66. Can the number of students be in fractions? This is absurd.
5. The arithmetic mean cannot be computed for open end classes. Since computation of A.M. for a continuous frequency distribution requires the class marks or mid points of class intervals, if there are open end class intervals it is not possible to find their mid points. Hence A.M. cannot be computed in such cases.

4.8 LET US SUM UP:

In this chapter we have learn:

- Meaning of measure of central tendency and its types.
- Calculation Arithmetic Mean (A.M.) for ungroup and group data.
- Calculation of weighted Mean.
- Calculate Combine Mean for two and more variables.

4.9 UNIT END EXERCISES:

1. What is the meaning of measure of central tendencies? Write down the characteristics of a good average.
2. Write a short note on different types of central tendencies.
3. Define Arithmetic Mean. What are the merits and demerits of A.M. Justify your answer with suitable Examples.
4. The daily income in '000 Rs. of 10 shopkeepers in Mumbai is as follows: 12, 08, 19, 25, 06, 30, 28, 41, 15, and 10. Find the average daily income.
5. The daily sales in numbers of vadapav in different areas in Mumbai is as follows:

Railway stn	Bus Depot	Schools	Colleges	Hospitals	Mantralaya
2245	2450	2500	1860	751	334

   Find the average number of vadapav's sold per day.
6. The marks obtained by 30 students in a test are given below. Find the average score of the class.

   10, 15, 22, 13, 26, 44, 23, 18, 10, 20, 36, 45, 29, 36, 11

   07, 39, 09, 33, 31, 25, 12, 41, 34, 26, 16, 18, 20, 31, 38

7. The profit (in lakhs of Rs.) of 10 companies in a city is as follows: 125, 224, 100, 435, 250, 565,

   280, 195, 320, 402. Find the average profit of all the companies.

8. The number of defective bulbs in 100 boxes is as follows: Find the average number of defective bulbs for all boxes.

Defective bulbs	0	1	2	3	4	5	6	7	8	9	10
Boxes	10	15	21	14	11	9	5	3	3	5	4

9. To mark the 150th anniversary of India's first freedom struggle of 1857, 120 students of a school were told to write down the martyrs they remember. The number of martyrs known to each student were as follows:

No. of martyrs	0	1	2	3	4	5	6	7	8	9	10
No. of students	12	38	24	11	16	07	06	03	01	02	00

10. The average weight of 30 students in a class is 40 kg. The following table gives the weight of 20

    students in the class. Find the average weight of remaining 10 students.

Weight (in kg)	34	36	38	40	42	44	46
Students	2	2	5	6	4	0	1

11. The following data shows the consumption of milk in litres in different families. Find the average consumption of milk.

Milk consumption	0.5	1	1.5	2	2.5	3
No. of families	10	28	30	21	16	5

12. The number of inquiries per day in a shoe shop for different sizes of shoes is given below. Find the average shoe size inquired about and what do you hence suggest the shopkeeper to do?

Size	5	6	7	8	9	10
No. of inquiries	14	32	28	25	10	5

13. The following table gives the marks obtained by students in the first term examination. Find the average marks per student

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	11	10	15	8	6

14. The following data gives the distance required by villagers of different villages of Thane district to come to the Collector's office. Find the average distance of travelling.

Distance (in km)	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of villages	98	62	35	60	152	70	189

15. The monthly wages of employees in a company is given below. Find the average wages.

Wages in '00 Rs.	5-15	15-25	25-35	35-45	45-55	55-65
No. of workers	05	16	22	18	10	09

16. The average life of two types of tube lights is given below. Find the average life of each type and comment on the result.

Life in months	0 – 6	6 – 12	12 – 18	18 – 24	24 – 30	30 – 36
Tube A	04	12	15	10	8	01
Tube B	06	10	14	08	05	07

17. The performance of students of three Universities A, B and C in different courses is given below. The data gives the percentage of students passed and no. of students (in'000). Using weighted average mean find out the best performing college.

Colleges	University A		University B		University C	
M.A.	55%	20	78%	40	80%	30
M.Com.	64%	30	70%	40	65%	20
M.Sc.	69%	10	65%	20	70%	50

18. Find the missing frequency from the following data if the total frequency is 100 and mean is 23.9

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	05	16	--	18	10	09

19. Find the missing frequency from the following data if the total frequency is 200 and mean is 49.75

Class Interval	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	25	--	42	51	36	17

20. Find the missing frequency if the mean for the following data is 9.725.

Income (in '000 Rs)	2 – 5	5 – 8	8 – 11	11 – 14	14 – 17	17 – 20
No. of families	14	22	8	?	12	04

21. If the average marks of 50 students in the tutorial test are 4.86, find the missing frequencies:

Marks	0	1	2	3	4	5	6	7	8	9	10
No. of students	2	6	?	11	4	3	?	4	2	7	5

22. The following frequency distribution gives number of students of a class who have passed an examination. Find their average marks. Also, find the average marks of the failed students if the total number students in the class were 100.

Marks	35-50	50-60	60-75	75-90	90-100
No. of students	12	15	10	11	12

23. The average salary of 400 employees in a private firm is Rs. 12,000. Due to increase in inflation, the firm decides to give an increment of 20% of the average salary to the highest paid employees, 15% of the average salary to the lowest paid employees and 10% of the average salary to the remaining employees. Find the extra payment the firm has to make for all employees. Also find the average salary after the increment.

24. The following table gives the salary per month of 200 employees in a company.

Salary	0-2500	2500-5000	5000-7500	7500-10000	10000-12500
No. of employees	50	40	45	35	30

The company offers an increment due to its profit in the first quarter as follows: 25% of the average salary to the highest paid employees, 20% of the average salary to the lowest paid employees and 15% of the average salary to the remaining employees. Find the extra payment the company has to make and also find the average salary per employee after the increment.

25. The average height of 100 students is 170 cm. The average height of 55 boys is 173 cm and that of the girls is 169 cm. Find the number of girl students.

26. The average pay of 25 men and 35 women in a factory are Rs. 100 and Rs. 80 respectively. Find the average pay of all the employees.

27. The average charge of a movie ticket for children and adults is Rs. 30 and Rs. 50 respectively. If 100 children and 150 adults watch a movie, what is the average charge of ticket of the audience?

28. The average marks of students in a class are 55. If the average marks for boys is 54 and the average marks for girls is 58. Find the ratio of

boys and girls in the class. If there are 60 boys in the class, find the number of girl students.

29. Mr. Vipul Patel owns three factories. The average wages for 50 labourers working in the first factory is Rs. 120, that of 80 labourers in second factory is Rs. 100 and for the 70 labourers in the third factory is Rs. 110. Find the average wage for all labourers in Mr. Patel's factories.
30. The average salary of 120 employees in a company is Rs. 12,000. The average salary of 20 Grade I employees is Rs. 16,000 and that of 40 Grade II employees is Rs. 12,400. Find the average salary of remaining employees.
31. The average marks of 20 students in a class are 75. If the average marks of 12 students are 70, find the average of remaining students in the class.
32. The mean height of 39 students in class is 164. The average becomes 164.2 because of the entry of a new student. What are the marks of the new student?
33. A salesman has average sales of Rs. 11,000 in the first 5 months of his job. Due to crash down in the market his sales for the sixth month are very low thereby decreasing the six monthly averages to Rs. 10,000. Find the sales made by the salesman in the sixth month.
34. The mean salary of 1000 employees in a company is Rs. 11,500. It was discovered that the salary of one employee was wrongly taken as 10,000 instead of 1,000. Find the correct mean salary.
35. The average weight of 50 people participating in a diet contest was calculated as 45 kg. But it was found that the actual weight of one of the participant was 62 kg and not 52 kg. Find the correct average weight of all participants.

Multiple Choice Questions:

1. If $\sum fx = 2120$ & $\sum f = 80$ then $\bar{x}$ is _____

   a) 26.5 b) 27.5 c) 37.5 d) 38.5
2. If there are two group with 100 observations each and 35 and 45 as values of their mean then the value of combined mean of 200 observations will be

   a) 35 b) 40 c) 45 d) None of these.
3. If n =5, $\sum xw$=1450, $\sum w$=80 then weighted mean is :-

   a) 40 b) 19.25 c) 36 d) 25
4. Mean or average used to measure central tendency is called

   a) sample mean. b) arithmetic mean.

   c) negative mean. d) population mean.
5. The mean of 25,15,20, 10, 30 is

   a) 25 b) 30 c) 20 d) 10

6. The arithmetic mean of asset of 10 numbers is 20. If each number is first multiplied by 2 and then increased by 5, then what is the mean of new numbers?

   a) 20 b) 25 c) 40 d) 45

7 What is the weighted mean of first 10 natural numbers whose weights are equal to the corresponding number?

   a) 7 b) 5.5 c) 5 d) 4.5

8. The arithmetic mean of first ten whole number is____.

   a) 5.5 b) 5 c) 4 d) 4.5

9. The mean of 5 observation is 25 of which first four observation are 35, 20, 40 and 5 than the fifth value is_____.

   a) 25 b) 30 c) 15 d) 50

10. The average weight of 50 people is 45 kg. If the average weight of 30 of them is 42 kg, than the average weight of remaining people is ___.

    a) 48.5 b) 50.5 c) 49.5 d) 51.5

4.10 LIST OF REFERENCES:

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

5

MEASURES OF CENTRAL TENDENCY II

Unit Structure

5.0 OBJECTIVES:

After going through this chapter you will able to know:

- Types of positional averages
- How to calculate central value using median.
- Extend median and calculate quartiles, Deciles, percentiles.
- Find median, quartile, deciles, percentiles by graphically using cumulative frequency curve.
- Calculate Mode for ungroup and ungroup data.
- Find mode graphically using histrogram.

5.1 INTRODUCTION

In previous chapter we have learn mathematical averages. Now in this chapter we are going to discuss about positional averages. The mean,

median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

5.2 MEDIAN

One of the limitations of arithmetic mean is that, it is affected by the extreme observations. To overcome this positional averages are very useful. Median is the first of the type of positional averages. It is the central value among a given set of observations written in either ascending or descending order of their magnitude.

5.2.1 MEDIAN FOR A UNGROUPED DATA:

Steps to find Median

1. First the given set of n observations is arranged in ascending or descending order.
2. If the number of observations (n) is odd, then median is the $\left(\frac{n+1}{2}\right)^{th}$ observation.
3. If the number of observations (n) is even, then the median is the arithmetic mean of $\left(\frac{n}{2}\right)^{th}$ observation and $\left(\frac{n}{2}+1\right)^{th}$ observation.

Example 1: Compute the median for the following series: 100, 78, 81, 43, 65, 77, 102, 34, and 59

Solution: We first arrange the given data in ascending order as follows

34, 43, 59, 65, 77, 78, 81, 100, 102

The numbers of observations are 9, *i.e.* odd in number.

Using the formula we have, Median = $\left(\frac{n+1}{2}\right)^{th}$ observation = $\left(\frac{9+1}{2}\right)^{\text{th}}$ observation

$\therefore$ Median = 5^{th} observation = 77.

Example 2: Compute the median for the following series: 9, 12, 17,8, 4, 15, 3, and 10

Solution: We first arrange the given data in ascending order as follows

3, 4, 8, 9, 10, 12, 15, 17

The numbers of observations are 8, *i.e.* even in number.

Now, $n/2 = 8/2 = 4$ and $n/2 + 1 = 5$

Using the formula Median = A.M. of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observation, we have,

$\therefore$ Median = A.M. of 4th and 5th observation = A.M. of 9 and 10 = 9.5

5.2.2 MEDIAN FOR A GROUPED (DISCRETE) DATA

Consider a set of observations $x_1, x_2, x_3 ....., x_n$ with corresponding frequencies $f_1, f_2, f_3 ....., f_n$.

Steps to find Median

1. First the less than cumulative frequencies are calculated.
2. We then find the total frequency N = $\sum f$.
3. Now we calculate the value of $N/2$. (Irrespective of whether N is odd or even.)
4. From the table we find the first cumulative frequency which is just greater than $N/2$. The corresponding observation is the median.

Example 3: The heights of 60 students in a class are given below. Find the median height.

Height in cm	165	166	167	168	169	170
No. of students	11	15	10	5	12	7

Solution: Introducing the column of less than cumulative frequencies we have,

Height in cm	No. of students (f)	Cumulative frequency
165	11	11($\not>$30)
166	15	26($\not>$ 30)
167	10	**36** (> 30)
168	5	41
169	12	53
170	7	60
Total	N = 60	

From the table N = 60. $\therefore$ N/2 = 30.

We look at the first cumulative frequency just greater than 30 in the third column. It is **36**

The corresponding value in the first column is 167.

$\therefore$ the Median = 167 cm.

5.2.3 MEDIAN FOR A CLASS DISTRIBUTION:

Consider a grouped and continuous class distribution with class intervals as $a_1 - a_2, a_2 - a_3, a_3 - a_4, \ldots etc$. Let the corresponding frequencies be denoted by $f_1, f_2, f_3, \ldots\ldots, f_n$.

Steps to find Median

1. First the less than cumulative frequencies are calculated.
2. We then find the total frequency N = $\sum f$.
3. Now we calculate the value of $N/2 = m$ (say).
4. The class interval whose cumulative frequency is just greater than $N/2$ is the median class where the median lies.
5. Let l_1 be the upper limit and l_2 be the lower limit of the median class. Let f denote the frequency of the median class and *pcf* denote the ***c***umulative ***f***requency of the ***p***revious class interval. Then the median is calculated by the formula:

   Median = $l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$, where $m = N/2$ and $i = l_2 - l_1$ is the width of the class interval.

Example 4: Find the median for the following data:

Class Interval	10-30	30-50	50-70	70-80	80-90	90-100
Frequency	17	34	20	52	36	23

Solution: Introducing the column of less than cumulative frequencies we have,

Class-Interval	Frequency	Cumulative frequency
10-30	17	17 ($\not>$ 91)
30-50	34	51 ($\not>$ 91)
50-70	20	**71** ($\not>$ 91)
70-80	**52**	123 (> 91)
80-90	36	159
90-100	23	182
Total	N = 182	

From the table N = 182. $\therefore m$ = N/2 = 91.

The first cumulative frequency just greater than 91 is 123. The corresponding class which is called as median class is 70 – 80.

Thus, $l_1 = 70$, $l_2 = 80$. $\therefore$ $i = l_2 - l_1 = 10$, $f = 52$ and $pcf = 71$

Now, Median $= l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right] = 70 + \left[\frac{(91-71)}{52} \text{ x } 10\right] = 70 + \left[\frac{20\text{x } 10}{52}\right]$

$\therefore$ Median $= 70 + 3.84 = 73.84$

Example 5: Find the median age for the following data:

Age	10-14	15-19	20-24	25-29	30-34	35-39
No. of persons	8	12	17	14	11	18

Solution: Since to compute median the class intervals need to be exclusive, we convert the given inclusive intervals to exclusive intervals by subtracting 0.5 from their lower limit and adding 0.5 to their upper limit. Introducing the column of less than cumulative frequencies we have,

Age	No. of Persons	Cumulative frequency
9.5–14.5	8	8 ($\not>$ 40)
14.5–19.5	12	20 ($\not>$ 40)
19.5–24.5	17	**37** ($\not>$ 40)
24.5–29.5	**14**	51 (> 40)
29.5–34.5	11	62
34.5–39.5	18	80
Total	N = 80	

From the table N = 80. $\therefore m$ = N/2 = 40.

The first cumulative frequency just greater than 40 is 51. The corresponding class which is called as median class is 24.5–29.5.

Thus, $l_1 = 24.5$, $l_2 = 29.5$. $\therefore$ $i = l_2 - l_1 = 5, f = 14$ and $pcf = 37$

Now, Median $= l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right] = 24.5 + \left[\frac{(40-37)}{14} \text{ x } 5\right] = 24.5 + \left[\frac{3 \text{ x } 5}{14}\right]$

$\therefore$ Median = 24.5 + 1.07 = 25.57.

5.2.4 GRAPHICAL METHOD FOR FINDING MEDIAN

The median can be found by graphical method using the following steps:

1. First the less than cumulative frequencies are calculated.
2. The upper limits of the class intervals are taken on the horizontal X-axis and the cumulative frequencies are taken on the vertical Y-axis.

3. Then we draw a less than ogive curve (cumulative frequency curve).

4. The value of $N/2$ is calculated and is marked on the Y-axis. Let this point be A.

5. From point A, a line parallel to X-axis is drawn till it touches the ogive curve at point B (say).

6. From B a line parallel to Y-axis is drawn till it touches the X-axis at point C (say). This point C is the required median.

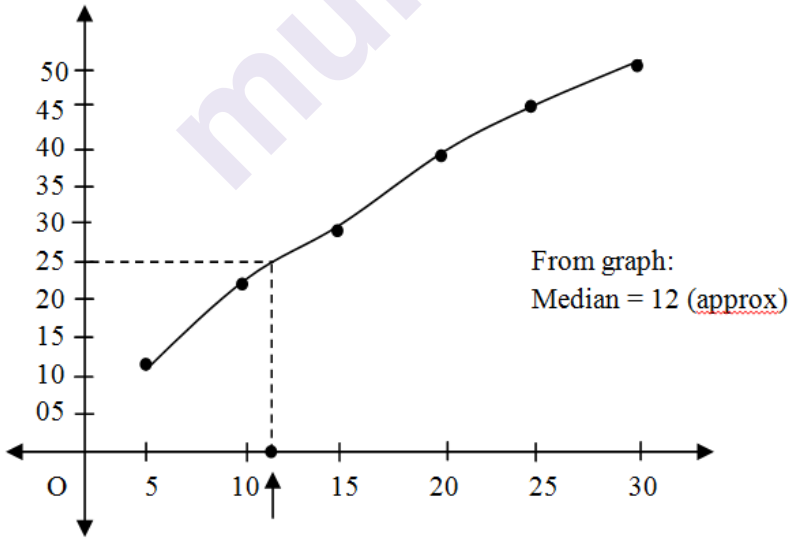
Example 6: Locate the median for the following data graphically:

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	52	36	42	38	54	44	34

Solution: Introducing the column of less than cumulative frequencies we have,

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	52	36	42	38	54	44	34
Cf	52	88	130	168	222	266	300

Now we plot the graph of upper limits of the class intervals taken on the horizontal X-axis against the cumulative frequencies taken on the vertical Y-axis.



Now, since N = 300, m = 150. We locate 150 on the Y-axis, draw a line parallel to X-axis till it touches the ogive curve and then drop a perpendicular on the X-axis. The point where it meets the X-axis is the median. In this case the approximate value of median is 35.

v Interested students can calculate median using formula and cross-check the graphical value of median.

Note: The graphical value of median is approximately same as the calculated value from formula provided a proper graph with right scale is drawn.

5.3 QUARTILES, DECILES AND PERCENTILES

The median is the value which divides the entire data in two equal parts. 50% of the observations are less than (or equal to) and 50% are greater than (or equal to) the median value. The value which divides the data in more than two parts is also useful for statistical calculations and analysis. A value which divides the entire data in four equal parts is called as Quartile. As we need three values to divide a set of values in four equal parts, there are three quartiles, namely first quartile Q_1, second quartile Q_2 and third quartile Q_3. Similarly, to divide a data in 10 equal parts we need 9 values called as Deciles. The *Deciles* are named D_1, D_2,…. D_9. Percentiles are those values which divide a data in 100 equal parts. There are 99 percentiles, *viz* P_1, P_2, …. , P_{99}.

QUARTILES

We know that Quartiles are the values which divide a set of observations in four equal parts. Q_1 is the value which has 25% observations less than or equal to it and 75% observations greater than or equal to it. For a continuous frequency distribution Q_1 is that value which represents 25% area under the histogram to the left of it. Q_2 is that value such that 50% observations are less than or equal to it and 50% observations are greater than or equal to it. In other words, Q_2 is nothing but Median. Q_3 is that value such that 75% observations are less than or equal to it and 25% observations are greater than or equal to it. For a continuous frequency distribution Q_3 is that value which represents 75% area under the histogram to the left of it and 25% of area under the histogram to right of it.

The algebraic formula and steps to calculate quartiles are similar to that of Median. The formulae are:

$$Q_1 = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right] \qquad Q_2 = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$$

$$Q_3 = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$$

Here $m = N/4$ Here $m = N/2$ Here $m = 3N/4$

The steps are exactly same with only difference in step 3 and step 4. The value of m depends on which quartile we are to find. As Q_1 divides the total frequency in the ratio of 1:4, m is taken as $N/4$. Q_3 divides the total frequency in the ratio 3:4, m is taken as $3N/4$.

Quartile, like median can be located graphically by taking the corresponding value of *m* on the Y-axis and proceeding in the same manner as stated in section **2.13**.

DECILES

Deciles are the values which divide the data into 10 equal parts. The steps for calculation and the formula for deciles are same as that of the quartiles with the difference of the value of *m*. For D_1, $m = N/10$, for D_2, $m = 2N/10 = N/5$, for D_3, $m = 3N/10$, and so on. In general, the formula for calculating the k^{th} decile is: $D_k = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$, where $m = \frac{kN}{10}$ and $k = 1, 2, 3, \ldots, 9$

PERCENTILES

Percentiles are the values which divide the data into 100 equal parts. The steps for calculation and the formula for percentiles are same as that of the quartiles and deciles with the difference of the value of *m*. For P_1, $m = N/100$, for P_2, $m = 2N/100 = N/50$, for P_3, $m = 3N/100$, and so on. In general, the formula for calculating the k^{th} percentile is:

$$P_k = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right], \text{ where } m = \frac{kN}{100} \text{ and } k = 1, 2, 3, \ldots\ldots, 99.$$

Note:

From the above formulae it is clear that $D_5 = Q_2$, $P_{25} = Q_1$, $P_{50} = Q_2 = D_5$, $P_{10} = D_1$, $P_{20} = D_2$ *etc.*

Example 7: Find the median, 1st & 3rd quartiles, 4th & 8th deciles and 30th & 60th percentile for the following data

Wages per day in Rs. :	50-100	100-150	150-200	200-250	250-300
No. of Workers :	10	24	39	65	52
Wages per day in Rs. :	300-350	350-400	400-450	450-500	500-550
No. of Workers :	45	34	26	15	14

Also find how many workers have wages between 175 and 375.

Solution: Introducing the less than cumulative frequency table:

Wages per day in Rs.	No. of workers (*f*)	Cumulative frequency
50-100	10	10
100-150	24	34
150-200	39	73

200-250	65	138
250-300	52	190
300-350	45	235
350-400	34	269
400-450	26	295
450-500	15	210
500-550	14	224

From the table N = 224.

(*i*) Median: $m = N/2 = 112$. The cumulative frequency just greater than 112 is 138. Thus, the median class is 200 – 250 and hence $l_1 = 200$, $l_2 = 250$. $\therefore\ i = l_2 - l_1 = 50$, $f = 65$ and $pcf = 73$

Using the formula for median, we have

$$\text{Median} = 200 + \left[\frac{(112-73)}{65} \text{ x } 50\right] = 230.$$

(*ii*) Quartiles: We have already calculated Q_2 which is the median.

For Q_1: $m = N/4 = 56$. The cumulative frequency just greater than 56 is 73. Thus, the first quartile class is 150 – 200 and hence $l_1 = 150$, $l_2 = 200$. $\therefore\ i = l_2 - l_1 = 50$, $f = 39$ and $pcf = 34$.

$$\therefore Q_1 = Q_1 = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right] = 150 + \left[\frac{(56-34)}{39} \text{ x } 50\right] = 178.20$$

For Q_3: $m = 3N/4 = 168$. The cumulative frequency just greater than 168 is 190. Thus, the third quartile class is 250 – 300 and hence $l_1 = 250$, $l_2 = 300$. $\therefore\ i = l_2 - l_1 = 50$, $f = 52$ and $pcf = 138$.

$$\therefore Q_1 = Q_1 = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right] = Q_3 = 250 + \left[\frac{(168-138)}{52} \text{ x } 50\right] = 278.84$$

(*iii*) Deciles:

For D_4: $m = 4N/10 = 89.6$ Proceeding in similar way we have $l_1 = 150$, $l_2 = 200$. $\therefore\ i = l_2 - l_1 = 50$

$f = 39$ and $pcf = 34$. Using the formula for 4th decile we get:

$$D_4 = 150 + \left[\frac{(89.6-34)}{39} \text{ x } 50\right] = 221.28$$

For D_{10}: m = 8N/10 = 179.2. Hence l_1 = 250, l_2 = 300. $\therefore$ $i = l_2 - l_1$ = 50, f = 52 and pcf = 138.

$$\therefore D_8 = 250 + \left[\frac{(179.2 - 138)}{52} \text{ x } 50\right] = 289.61$$

(*iv*) Percentiles:

For P_{30}: m = 30N/100 = 67.2. Hence l_1 = 150, l_2 = 200. $\therefore$ $i = l_2 - l_1$ = 50, f = 39 and pcf = 34.

$$\therefore P_{30} = 150 + \left[\frac{(67.2 - 34)}{39} \text{ x } 50\right] = 192.56$$

For P_{60}: m = 60N/100 = 134.4. Hence l_1 = 200, l_2 = 250. $\therefore$ $i = l_2 - l_1$ = 50, f = 65 and pcf = 73

$$\therefore P_{60} = 200 + \left[\frac{(134.4 - 73)}{65} \text{ x } 50\right] = 247.23$$

(*v*) To find the number of workers whose wages are between 175 and 375.

Clearly the class intervals 200 – 250 and 250 – 300 include the number of workers required, which is 65 + 52 = 117.

The number of workers with wages between 175 and 200 is calculated as :

$$\frac{200 - 175}{50} \text{ x } 39 \approx 20$$

The number of workers with wages between 350 and 375 is calculated as :

$$\frac{375 - 350}{50} \text{ x } 34 = 17$$

Thus, the number of workers whose daily wages are between Rs. 175 and Rs. 375

= 117 + 20 + 17 = 154.

Note: In the above problem, the calculation of no. of workers with wages between a certain limit is done using the following formula: $\frac{|l - x|}{i} \text{ x } f$, where l: class limit (upper or lower) of interval,

x: given observation, i: width if the interval and f: frequency of that class interval.

For example: In the interval 20 – 40 with corresponding frequency 15, if we want to know how many observations are there above 30, then using above formula,

$$\text{no. of observations} = \frac{40 - 30}{20} \text{ x } 15 = 7.5$$

In the interval 10 – 20 with corresponding frequency 12, if we want to know how many observations are there below 15, then using above formula, no. of observations = $\frac{15-10}{20}$ x $12 = 3$

Example 8: The following data gives the scores of 150 candidates who appeared for a test. Find the 6th decile and 66th percentile for the data:

Scores	50-55	55-60	60-65	65-70	70-75	75-80	80-85	85-90	90-95	95-100
Candidates	12	15	14	21	17	18	20	18	7	8

Solution: Introducing the column of less than cumulative frequency we have,

Scores	50-55	55-60	60-65	65-70	70-75	75-80	80-85	85-90	90-95	95-100
Candidates	12	15	14	21	17	18	20	18	7	8
cf	12	27	41	62	79	97	117	135	142	150

From the table: $N = 150$.

(*i*) The formula for D_6 is: $D_6 = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$, where $m = 6N/10 = 3N/5$

$\therefore\ m = 90$. Hence $l_1 = 75$, $l_2 = 80$, $i = 80 - 75 = 5$, $f = 18$ and $pcf = 79$.

$\therefore D_6 = 75 + \left[\frac{(90-79)}{18} \text{ x } 5\right] = 78.05$

(*ii*) The formula for 66th Percentile is : $P_6 = l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$, where $m = 66N/100 = 33N/50$

$\therefore\ m = 99$. Hence $l_1 = 80$, $l_2 = 85$, $i = 85 - 80 = 5$, $f = 20$ and $pcf = 97$.

$\therefore P_6 = 80 + \left[\frac{(99-97)}{20} \text{ x } 5\right] = 80.5$

MISSING FREQUENCY PROBLEMS

Example 9: Find the missing frequency in the following data with median = 27.5

Profit	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
Firms	4	--	20	10	7	3

Solution: Let the missing frequency be k. Preparing the cumulative frequency distribution table we have:

Profit in lakhs of Rs.	Firms	*Cf*
0 – 10	4	4
10 – 20	k	**4 + *k***
20 – 30	**20**	24 + k
30 – 40	10	34 + k
40 – 50	7	41 + k
50 – 60	3	44 + k
Total	N = 44 + k	

Since the median = 27.5, it means that the median class is 20 – 30.

Hence, $m = \frac{N}{2} = \frac{44+k}{2}$, $l_1 = 20$, $l_2 = 30$, $i = 30 - 20 = 10$, $f = 20$ and $pcf = 4 + k$

Using the median formula : Median = $l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$, we have

$$27.5 = 20 + \left[\frac{\left(\frac{44+k}{2} - (4+k)\right)}{20} \text{ x } 10\right] = 20 + \left[\frac{(44+k) - 2(4+k)}{2 \text{ x } 20} \text{ x } 10\right]$$

$$\therefore 27.5 = 20 + \left[\frac{44+k-8-2k}{4}\right] \Rightarrow 27.5 - 20 = \frac{36-k}{4}$$

$$\therefore 7.5 = \frac{36-k}{4} \Rightarrow 30 = 36 - k$$

$\therefore k = 6$. The missing frequency is 6.

Example 10: If the total frequency for the following data is 100 and median is 31.5 find the missing frequencies.

C.I.	0.5–10.5	10.5–20.5	20.5–30.5	30.5–40.5	40.5–50.5	50.5–60.5	60.5–70.5
Frequency	8	-	25	20	16	-	6

Solution: Let the two missing frequencies be a and b. Preparing the frequency distribution table we have:

C.I.	Frequency	*cf*
0.5–10.5	8	8
10.5–20.5	*a*	8 + *a*
20.5–30.5	25	33 + *a*
30.5–40.5	20	53 + *a*
40.5–50.5	16	69 + *a*
50.5–60.5	*b*	69 + *a* + *b*
60.5–70.5	6	75 + *a* + *b*

Now, from the table $N = 75 + a + b$. But it's given that $N = 100$

$\therefore 75 + a + b = 100$

$\Rightarrow a + b = 25 \quad \ldots (*)$

Since median = 31.5, the median class is 30.5 – 40.5.

Hence, $m = N/2 = 100/2 = 50$, $l_1 = 30.5$, $l_2 = 40.5$, $i = 40.5 - 30.5 = 10$, $f = 20$ and $pcf = 33 + a$

Using the median formula : Median $= l_1 + \left[\frac{(m - pcf)}{f} \text{ x } i\right]$, we have

$$31.5 = 30.5 + \left[\frac{50 - (33 + a)}{20} \text{ x } 10\right]$$

$$\therefore 31.5 - 30.5 = \frac{50 - 33 - a}{2} \Rightarrow 1 = \frac{17 - a}{2} \Rightarrow 2 = 17 - a$$

$\therefore a = 15 \qquad (1)$

From (*): $a + b = 25$ and from (1): $a = 15$

$\therefore b = 10 \qquad (2)$

Thus, the missing frequencies are 15 and 10 respectively.

5.5 MERITS AND DEMERITS OF MEDIAN

Merits:

1. It is rigidly defined.
2. It can be easily calculated and also understood.
3. It can be calculated even if some extreme observations are incomplete.
4. It is not affected by extreme observations in the data.

5. It can be located graphically using ogive curves.
6. It gives the justice to find the average of a qualitative attribute of the data.

Demerits

1. It is difficult to arrange large number of data in ascending or descending order.
2. It is not useful for any further algebraic treatment.
3. Since it is not affected by the extreme values, it may not be a true representative of a data where the extreme values are important.
4. It is likely to be affected by the sampling variations.

5.5 MODE:

Many a times it is important to know the most likely value in a set of data. For Example, if we need to know, which is the most commonly read book or newspaper in a library or the most common eatable at a stall *etc.* Mode is that measure of central tendency which gives a number representing the most likely item. It is denoted by *Z*.

5.5.1 Mode for ungroup data:

For a raw data, it is defined as the observation which occurs maximum number of times among a set of observations.

Example 11: Find the mode for the following data: 10, 12,11,15,12, 11, 16, 19, 12, 10, 19, 12, 15, 20, and 12.

Solution: By mere inspection we can see that the value 12 is repeated maximum number of times *i.e.* 5 times and hence the mode for the given data is 12.

Note: 1. If all value are identical than mode is same value.

2. If all value are distinct than the mode does not exists.

5.5.2 Mode for group data:

For a **discrete data**, the observation with highest frequency is defined as Mode.

Example 12: Calculate the mode for the following distribution:

X	5	6	7	8	9
F	12	27	35	23	14

Solution: Here maximum frequency is 35.

Maximum frequency is corresponding to the group 7.

Therefore, Mode = 7.

For a **continuous frequency distribution**, mode is calculated by the following steps:

1. Let f_1 : Highest frequency and l_1 & l_2 denote the lower and upper class limits of the corresponding modal class.
2. Let f_0 : frequency of pre-modal class and f_2 : frequency of post-modal class.
3. The Mode is now calculated by the formula: $Z = l_1 + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \text{ x } i\right]$, where $i = l_2 - l_1$

   This formula can also be stated as $Z = l_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \text{ x } i\right]$, where $\Delta_1 = f_1 - f_0$ and $\Delta_2 = f_1 - f_2$.

Example 13: The following data gives the ages of the number of drop out students in a village. Find the modal age.

Age group	3-5	5-7	7-9	9-11	11-13	13-15	15-17
Frequency	25 **(f_0)**	71 **(f_1)**	44 **(f_2)**	33	58	30	39

Solution: This is a continuous frequency distribution. Since the maximum frequency is 71, the modal class is 5 – 7. Hence, $l_1 = 5$, $l_2 = 7$, $i = 7 - 5 = 2$, $f_0 = 25$, $f_1 = 71$ and $f_2 = 44$. $\therefore \Delta_1 = f_1 - f_0 = 71 - 25 = 46$ and $\Delta_2 = f_1 - f_2 = 71 - 44 = 27$.

Now, the formula to calculate mode now is: $Z = l_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \text{ x } i\right]$

$\therefore Z = 5 + \left[\frac{46}{46 + 27} \text{ x } 2\right] \Rightarrow Z = 5 + 1.26 = 6.26.$

Example 14: The following data gives the population of women of different age groups in a city. Find the modal age

Age group	1-10	11-20	21-30	31-40	41-50	51-60	61-70
No. of women	1022	1239 **(f_0)**	1350 **(f_1)**	768 **(f_2)**	981	1074	739

Solution: This is inclusive type of data. First the age groups are converted to exclusive type by subtracting and adding 0.5 to the lower and upper class limits respectively.

Now the maximum frequency is 1350, the modal class is 20.5 – 30.5. Hence, $l_1 = 20.5$, $l_2 = 30.5$,

$i = 30.5 - 20.5 = 10$, $f_0 = 1239$, $f_1 = 1350$ and $f_2 = 768$. $\therefore \Delta_1 = f_1 - f_0 = 1350 - 1239 = 111$ and

$\Delta_2 = f_1 - f_2 = 1350 - 768 = 582$.

Now, the formula to calculate mode now is: $Z = l_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \text{ x } i\right]$

$\therefore Z = 20.5 + \left[\frac{111}{111+582} \text{ x}10\right] = 20.5 + 1.6 = 22.1 \approx 22$

$\therefore$ the modal age is 22 years.

Missing frequency problems involving mean, median and mode

Example 15: If the median is 27.41 and Mode is 25.63 for the following data, find the missing frequencies.

C.I.	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
Frequency	12	--	27	20	--	6

Solution: Let the missing frequencies be a and b.

C.I.	Frequency	*Cf*
0 – 10	12	12
10 – 20	a	$12 + a$
20 – 30	27	$39 + a$
30 – 40	20	$59 + a$
40 – 50	b	$59 + a + b$
50 – 60	6	$65 + a + b$

Given $Z = 25.63 \Rightarrow$ the modal class is 20 – 30. Now, $l_1 = 20$, $l_2 = 30$, $i = 30 - 20 = 10$, $f_0 = a$, $f_1 = 27$, and $f_2 = 20$. $\therefore \Delta_1 = f_1 - f_0 = 27 - a$ and $\Delta_2 = f_1 - f_2 = 27 - 20 = 7$.

Now, the formula to calculate mode now is: $Z = l_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \text{ x } i\right]$

$\therefore 25.63 = 20 + \left[\frac{27-a}{27-a+7} \text{ x}10\right] \Rightarrow 25.63 - 20 = \frac{270-10a}{34-a}$

$\therefore 5.63(34-a) = 270 - 10a \Rightarrow 191.42 - 5.63a = 270 - 10a$

$\therefore 4.37a = 78.58 \Rightarrow a = 17.98 \approx 18$ (1)

Now from table and (1) we have: $N = 65 + a + b = 65 + 18 + b = 83 + b$

Since Median = 27.41, median class is again 20 – 30. Hence, $m = N/2 = \frac{83+b}{2}$, $l_1 = 20$, $l_2 = 30$,

$f = 27$ and $pcf = 12 + a = 12 + 18 = 30$ …from (1)

$$\text{Median} = l_1 + \left[\frac{m - pcf}{f} \text{ x } i\right] \qquad \therefore 27.41 = 20 + \left[\frac{\frac{83+b}{2} - 30}{27} \text{ x } 10\right]$$

$$\therefore 27.41 - 20 = \left[\frac{83+b-60}{27 \text{ x } 2} \text{ x } 10\right] \Rightarrow 7.41 = \frac{230+10b}{54}$$

$$\therefore 400.14 = 230 + 10b \Rightarrow 170.14 = 10b \Rightarrow b = 17.01 \approx 17 \qquad (2)$$

Thus the missing frequencies are 18 and 17.

5.5.3 GRAPHICAL LOCATION OF MODE

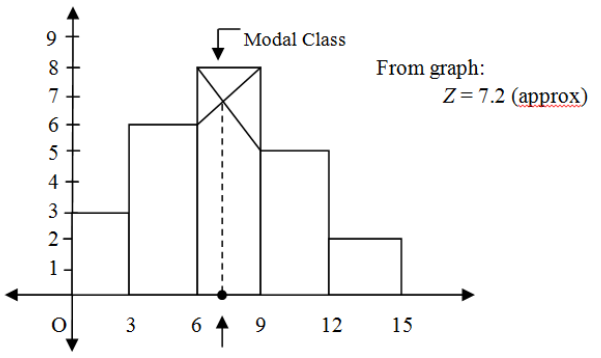
Mode for a grouped frequency distribution can be computed graphically by drawing the Histogram representing the data. The steps involved are as follows:

1. The Histogram representing the data is drawn.
2. Two diagonals are drawn connecting the upper corners of the bar representing the modal class to the upper corners of the adjacent bars.
3. A perpendicular is drawn from the point of intersection of these two diagonals to meet the X-axis at a point say *Z*. This value is the Mode of the distribution.

Example 16: Locate the mode graphically for the following data:

Class Interval	0 – 3	3 – 6	6 – 9	9 – 12	12 – 15
Frequency	3	6	8	5	2

Solution: The Histogram representing the above frequency distribution is drawn as follows:



5.6 MERITS AND DEMERITS OF MODE:

Merits

1. It is simply defined and hence is easy to understand and calculate.
2. It can be easily located from the graph.
3. It is not affected by the extreme values of the data.
4. It may be used for describing quantitative and qualitative data.
5. Because of its most likely approach, mode is a popular average for common people to areas like Biostatistics.

Demerits

1. In case of bimodal or multimodal frequency distributions it is not possible to interpret or compare.
2. It is not rigidly defined. Different methods may give different answers
3. It is not based on all observations and hence does not represent the entire set of data.
4. It cannot be used for further algebraic treatment.

5.7 COMPARATIVE ANALYSIS OF ALL MEASURES OF CENTRAL TENDENCY :

For a given data, a choice of an average is very important. Any choice should be made by taking into consideration the following:

(*i*) Purpose of computing an average.

(*ii*) Need for further treatment of the average.

(*iii*) Type of data in hand.

(*iv*) Merits and Demerits of an average over the other.

Arithmetic Mean can be generally used for all purposes.

Geometric Mean can be used where the small observations require more weight and vice versa.

Median being a positional average can be used for qualitative interpretation or when the data is incomplete.

Mode can be used when the purpose is to find the most likely type of value from the data.

Finally we conclude with some funny yet important note. Drawing conclusions from any statistical measure is no doubt the important job but also a difficult one. If the averages are not interpreted properly it may lead to absurd conclusions. Let us see some Examples:

(*i*) The average depth of a swimming pool is 5ft.

Wrong Conclusion: Any person of height more than 5 ft for e.g. 6ft can cross the pool easily!

(*ii*) The average marks of students of Division *A* of FYBMS class of a college are 20% and that of Division *B* are 50%.

Wrong Conclusion: All students of Division *A* are not academically strong as compared to those of Division *B*!

(*iii*) On an average, only 10% casualties related to local trains are due to travelling on the roof, while the other reasons constitute 90% casualties.

Wrong Conclusion: It is safer to travel on the roof of a local train!

(*iv*) The average daily income of a resident of Mumbai is Rs. 5,000.

Wrong Conclusion: All Mumbaikars, from beggars (well that may be true!!) to entrepreneurs earn more than or at least Rs. 5,000 per day.

(*v*) In a society with 10 flats, the number of children in each family is 2, 1, 3, 2, 2, 1,0, 2, 3, 2.

Wrong Conclusion: (*a*) Every family has at least 2 children! This is based on Mode.

(*b*) Every family has 1.8 children! This is based on arithmetic mean.

(Don't ask where to bring that 0.8^{th} child from!!)

RELATION BETWEEN MODE FROM MEAN AND MEDIAN

Mode can also be calculated using mean and median by the experimental approximate formula given by Karl Pearson:

Mode = 3Median – 2Mean

If the values of mean and median for a certain are known, then the mode can be calculated by the above formula.

Example 17: If the mean and median of a certain data is 34.6 and 38 respectively, find the mode

Solution: Using the Karl Pearson's formula, we have

Mode = 3Median – 2Mean

$\therefore Z = 3(38)-2(34.6) = 44.8$

5.8 LET US SUM UP:

In this chapter we have learn:

- Types of positional averages.
- To calculate median for ungroup and group data.
- To calculate median graphically.
- To calculate quartiles, deciles and percentiles.
- To calculate mode foe ungroup and group data.
- To calculate mode graphically.

- To find relation between mean, median and mode.

5.9 UNIT END EXERCISES:

1. Define Mean, Median and Mode. What are the advantages of median over mean and mode over mean?
2. State giving examples, the factors based on which a measure of central tendency is selected.
3. What are important features of a good average?
4. Discuss with examples the difference between a simple arithmetic mean and weighted mean.
5. Write a short note on graphical representations of averages.
6. The heights in cm of 12 boys in a class are given as follows: 154, 157, 162, 158, 171, 169, 153, 156, 157, 164, 166, 170. Find the median age.
7. The speed of a bus at 10 check points was observed as follows: 40, 35, 42, 50, 40, 60, 45, 50, 44, 52. Find the median speed of the bus.
8. The weights in kg of 13 children are as follows: 36, 32, 40, 39,35, 43, 41, 38, 30, 29, 37, 33, 40. Find the median weight.
9. The number of telephone calls made every successive hour from a PCO is given below. Find the median.

No. of calls	10	15	20	25	30	35
Frequency	26	30	16	22	10	6

10. The IQ test of 145 students is as follows: Find the media, 1st quartile, 8th decile and 24th percentile.

C.I.	50-70	70-90	90-110	110-130	130-150	150-170	170-190
No. of students	22	38	40	10	15	5	20

11. Locate the quartiles graphically for the following data:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of students	37	55	62	30	28	39	40	100	67	42

12. If the median for the following distribution is 17.5, find the missing frequency.

C.I.	0-10	10-20	20-30	30-40	40-50
Frequency	10	?	20	00	10

13. If the median and mode for the following data is 27.5 and 25.83 (approx), find the missing frequencies.

Marks (less than)	10	20	30	40	50	60
No. of students	4	10	30	40	47	50

14. The weights of Alphanso mangoes which are to be exported are given below. The standard weight for export quality mango is 800 gm. What percentage of the total mangoes will qualify to get exported?

Weights	200-400	400-600	600-700	700-800	800-900	900-1000	1000-1100	1100-1200
No. of students	37	55	62	30	28	39	40	100

15. Find the mode for the following data: 11, 23, 12, 11, 32, 23, 18, 11, 22, 24, 31, 11, 15, 18, 12, 11.

16. In an examination, the questions attempted by 100 students are given below. Find the mode.

Q. No.	1	2	3	4	5	6	7	8
No. of attempts	45	58	77	35	49	70	22	48

17. If the mode for the following data is 55, find the missing frequency for the following data:

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	3	5	7	10	?	15	12	6	2	8

18. If the mean and median values are 67.56 and 61.23, find the modal value.

19. If the average marks of 120 students are 75 and their modal marks are 70, find the median marks.

20. Locate the mode for the following data graphically.

C.I.	0-10	0-20	0-30	0-40	0-50	0-60	0-70	0-80	0-90	0-100
Frequency	6	18	33	47	60	80	92	100	115	120

21. Locate the mode graphically for the following data:;

Age	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40
No. of persons	14	28	40	30	22	16

22. Calculate the mean, median and mode for the following data:

Income in '000 Rs.	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
No. of families	13	15	20	22	16	10	08	05	09	02

23. Calculate the mean, median and mode for the following data:

C.I.	5-15	15-25	25-35	35-45	45-55	55-65	65-75	75-85	85-95
Frequency	4	12	16	20	14	18	03	12	01

24. If the mean and median of the following data is 23 and 22.77, find the missing frequency.

CI	0-10	10-20	20-30	30-40	40-50
Frequency	16	?	18	?	10

Multiple Choice Questions:

1. Which one divide the data in four equal parts :-

   a) Mode b) Median c) Quartile d) Decile

2. In case of extreme values the best measure of central tendency is :-

   a) A.M. b) Median c)Mode d) None of the above

3. _____the value of the middle observation when the observation are arranged in the order of their magnitude .

   a) Mean b) Median c) Standard deviation

   d) Variance

4. The mode of the data, 1, 2, 2, 2, 3, 3, 1,5 is_____

   a) 1 b) 2 c) 3 d) 5

5. Percentiles divide the data into_____

   a) 100 b) 200 c) 300 d) 110

6. The mode of the following data, is_____

X	12	14	18	20	22
F	2	18	24	12	8

a) 18 b) 22 c) 24 d) 8

7. Percentiles divide the data into 100 parts using _____ number of values.

   a) 100 b) 99 c) 90 d) 101

8. In deciles, central tendency median to be measured must lie in

   a) fourth deciles. b) seventh deciles c) eighth deciles. d) fifth deciles.

9. When data is arranged, middle value in set of observations is classified as

   a) median. b) mean. c) variance. d) standard deviation.

10. In a negative skewed distribution, order of mean, median and mode is as

    a) mean < median > mode. b) mean > median > mode.

    c) mean < median < mode.m d) mean > median < mode.

5.10 LIST OF REFERENCES:

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

6

MEASURE OF DISPERSION

Unit Structure

6.0 OBJECTIVES:

After going through this chapter you will able to know:

- Meaning and function of measure of dispersion.
- Types of measure of dispersion.
- Different method of calculating measure of dispersion.
- Concept of Skewness and Kurtosis.

6.1 INTRODUCTION:

In the previous chapter we have seen that how a measure of central tendency represents the entire data. Though an average is of great importance for statistical analysis, it has its limitations, as seen in the last section of chapter

two. There can be cases wherein the averages may come out to be same but the individual observations and their trends may be completely different. For example, let us consider the following three sets of data:

Data I :0, 10, 30, 50, 80, 130

Data II :45, 57, 48, 59, 60, 31

Data III: 50.2, 49.8, 48.7, 50.5, 50.7, 50.1

In all the three cases, the average is 50. But is it correct hence, to conclude all the three data show a similar trend of observations? The answer is no. The first type of data is completely away from the central value 50. In the next case, the data is scattered around the central value. While in the third case, the data is densely scattered around the central value. So, if we analyse the given data only on the basis of its average then we may not get a right conclusion. There is something more which needs to be known. There are cases when we may be interested more in how is the trend of the observations and not its average. In view of the above example, we may be interested in knowing how much is the data scattered from the central value. This extent of scatter is called as dispersion. The lesser the dispersion it means that the average is a true representative of the entire data. More dispersion indicates that the average is not the true representation of the data. This information is most important for testing of hypothesis, testing the consistency or for forecasting purposes in statistical analysis.

Dispersion is measured as an average of the deviations of all the observations from the central value. We know that the measures of central tendencies are referred as the averages of first order. Since dispersion is calculated from the central value or averages, it is often called as average of second order.

6.2 FUNCTION OF MEASURES OF DISPERSION:

The following are the functions of a measure of dispersion:

1. To test the reliability of the measures of central tendencies

   The measures of dispersion help in testing whether an average is a true representation of the data. If the dispersion is small, then it means that the observations are scattered around the average value. In other words, the average taken is reliable. If the dispersion is high, then it means that the observations are scattered away from the central value and hence, the average taken in not reliable. Zero dispersion indicates that all the observations are identical.

2. To facilitate in identifying and rectifying the causes of variations

   Measures of dispersion facilitates in identifying the extent of variations of the observations from the average value. This information may be used to rectify and take proper measure to control the variations. In areas of Biostatistics and Sociology it is used to understand the nature and extent of the causes of deviations from the average values.

3. To compare the variability of two or more series

   The most important function of a measure of dispersion is its utility to compare the variations in different sets of data. The consistency of players, performance of students, insurance policies or its agents, share prices of different firms *etc* can be calculated using measures of dispersion.

4. To help in computations for further statistical measures

   The measures of dispersion are used for further statistical measures like correlation anlaysis, regression analysis, forecasting, testing of hypothesis, analysis of variance *etc*.

6.3 FUNDAMENTALS OF DISPERSION:

The fundamentals of a measure of dispersion are same as that of the measures of central tendencies. The properties are as follows:

- It should be easy for calculation.
- It should be simple to understand.
- It should be rigidly defined.
- It should not be affected by the extreme observations.
- It should be based on all observations.
- It should not be affected by sampling fluctuations.
- It should be available for further algebraic treatment.

6.4 TYPES OF DISPERSION:

The measures of dispersion are classified in two types: (1) Absolute measure and (2) Relative measure.

(1) **Absolute measure:** The measure of dispersion which is expressed as the absolute variation between the observations from the central value, in the same units that of the observations, is called as an absolute measure. This is useful in comparing the variations in two or more sets of data measured in same units.

(2) **Relative measure:** The measure of dispersion which is expressed as the ratio of the absolute deviations to the central value is called as relative measure.

The different absolute and relative measures of dispersion are:

Absolute measure	Relative measure
Range	Coefficient of Range
Quartile Deviation	Coefficient of Quartile Deviation

Mean Deviation (from mean, median or mode)	Coefficient of Mean Deviation
Standard Deviation	Coefficient of Variation

6.5 RANGE:

It is one of the simplest and elementary measures of dispersion. Range is the difference between the smallest and the largest observation of a data.

Range of an ungrouped data:

Symbolically, if L denotes the **l**argest observation and S denotes the **s**mallest observation, then the absolute measure for an ungrouped data, Range denoted by R is given by: $R = L - S$.

For example, if the height in cm of 6 students in a class are 156, 160, 145, 161, 167, 164 then the $R = 167 - 145 = 22$ cm. This gives the variability in the height of students.

Range of a grouped data

If the data is grouped into class intervals then the range can be computed by any of the following methods:

(1) $R = U - L$, where $U :=$ upper limit of the highest class interval and $L :=$ lower limit of the lowest class interval.

OR

(2) $R = x_U - x_L$, where $x_U :=$ mid point of the highest class interval and $x_L :=$ mid point of the lowest class interval.

6.5.1 Coefficient of range:

As in the above example if the weight in kg of the same 6 students is 42, 39, 51, 47, 55, 40 then the range is $R = 55 - 39 = 16$ kg. But both the range values cannot be compared as one of them is in cm and the other is in kg. Thus, a relative measure which independent of the unit of observations is used called as *coefficient of range*.

The relative measure of range *i.e.*, *Coefficient of Range* $= \frac{L-S}{L+S}$ or $\frac{U-L}{U+L}$

Example 1: Find the range for the following data giving the distances covered in km by different missiles. 1000, 600, 3000, 3500, 2000, 1500, 1200 and 250. Also compute coefficient of range.

Solution: Here $L = 3500$ and $S = 250$

$\therefore$ Range $R = L - S = 3500 - 250 = 3250$ km.

Coefficient of Range $= \frac{L-S}{L+S} = \frac{3500-250}{3500+250} = \frac{3250}{3750} = 0.87$

Example 2: Find the range for the following data giving marks of 50 students in a class test of 50 marks. Also find the coefficient of range.

Marks	0 – 10	10 – 20	20 – 30	30 – 40
No. of students	08	15	17	10

Solution: Here the upper limit of the highest class interval 30 – 40 is U = 40 and the lower limit of the lowest class interval 0 – 10 is L = 0.

$\therefore R = U - L = 40 - 0 = 40.$

Coefficient of Range $= \frac{U-L}{U+L} = \frac{40-0}{40+0} = 1.$

Example 3: The coefficient of range of a certain set of observations is 0.6. If the smallest observation is 100, find the largest observation.

Solution: Given S = 100 and coefficient of range *i.e* $\frac{L-S}{L+S} = 0.6$

$\therefore \frac{L-100}{L+100} = 0.6 \quad \Rightarrow L - 100 = 0.6(L+100)$

$\therefore L - 100 = 0.6L + 600 \quad \Rightarrow L - 0.6L = 100 + 600$

$\therefore 0.4L = 700 \Rightarrow L = 1750$

Thus, the largest observation is 1750.

6.5.2 Merits and demerits of range:

Merits:

(1) It is easy to compute.

(2) It is easy to understand.

Demerits:

1. **It is not based on all observations:** Range is computed only on the basis of extreme values. So, two different types of data with same extreme values will have same dispersion, eventhough the individual sets of data may be having completely different observations. Thus, it does not measure the dispersion of all observations.

2. **It is easily affected by the extreme observations:** Again, it is obviously affected by the change in extreme values.

3. **It is readily affected by sampling fluctuations:** If the samples taken include or exclude the extreme values of the entire population, the value of the range will be completely diverse.

4. **It cannot be used for open-end frequency distributions:** If a given data has unknown lower class limit of the lowest class or unknown upper class limit if the highest class, it is not possible to compute range.

As can be seen clearly, *Range* has more demerits to its credit than merits; still it is a popular measure and is used widely in certain fields.

In data related to small fluctuations, range is commonly used. For example, to study the temperature fluctuations of a city in day time, share prices of certain firm, rainfall, prices of commodities, currency rate, to prepare control charts and also as a quality control measure.

6.6 SEMI – INTER QUARTILE RANGE OR QUARTILE DEVIATION (Q.D.):

Semi – inter quartile range also called as quartile deviation, is the mid point of the inter quartile range. Symbolically, Q.D. = $\frac{Q_3 - Q_1}{2}$. It is an absolute measure of dispersion.

6.6.1 Coefficient of quartile deviation:

The corresponding relative measure of Q.D. is defined as follows:

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

Example 4: Find the Quartile Deviation of the daily wages (in Rs.) of 11 workers given as follows: 125, 75, 80, 50, 60, 40, 50, 100, 85, 90, 45.

Solution: Arranging the data in ascending order we have the wages of the 11 workers as follows:

40, 45, 50, 50, 60, 75, 80, 85, 90, 100, 125

Since the number of observations is odd (11), the 1st Quartile is given by:

$Q_1 = (11 + 1)/4 = 3^{rd}$ observation = 50. … (1)

$Q_3 = 3(11 + 1)/4 = 9^{th}$ observation = 90. … (2)

Q.D. = $\frac{Q_3 - Q_1}{2}$ = (90 – 50)/2 = 20 … from (1) and (2)

Example 5: The following data gives the weight of 60 students in a class. Find the range of the weights of central 50% students.

Weight in kg	30 – 35	35 – 40	40 – 45	45 – 50	50 – 55	55 – 60
No. of students	4	16	12	8	10	5

Solution: To find range of the weight's of central 50 % students means to find the inter quartile range. For that we require Q_1 and Q_3. The column of less than *c f* is introduced as follows:

Weight in kg	30 – 35	35 – 40	40 – 45	45 – 50	50 – 55	55 – 60
No. of students	4	**16**	12	8	**10**	6
cf	**4**	20	32	**40**	50	56
		↑ Q_1 class			↑ Q_3 class	

To find Q_1: $N = 56$. Thus $m = N/4 = 14$.

The *cf* just greater than 14 is 20, so 35 – 40 is the 1st quartile class and

$l_1 = 35$, $l_2 = 40$, $i = 40 - 35 = 5$, $f = 16$ and $pcf = 4$.

$$\therefore Q_1 = l_1 + \left[\frac{m - pcf}{f} \text{ x } i\right] = 35 + \left[\frac{14-4}{16} \text{ x } 5\right] = 35 + 3.125 = 38.125$$

To find Q_3: $N = 56$. Thus $m = 3N/4 = 42$.

The *cf* just greater than 42 is 50, so 50 – 55 is the 3rd quartile class and

$l_1 = 50$, $l_2 = 55$, $i = 55 - 50 = 5$, $f = 10$ and $pcf = 40$.

$$\therefore Q_1 = l_1 + \left[\frac{m - pcf}{f} \text{ x } i\right] = 50 + \left[\frac{42-40}{10} \text{ x } 5\right] = 50 + 1 = 51$$

$\therefore$ inter quartile range = $Q_3 - Q_1 = 51 - 38.125 = 12.875$ kg

Thus, the range of weight for the central 50% students = 12.875kg

Example 6: Find the semi – inter quartile range and its coefficient for the following data:

Size of shoe	0	1	2	3	4	5	6	7	8	9	10
No. of boys	7	10	15	11	18	10	16	5	12	6	2

Solution: The less than *c f* are computed and the table is completed as follows:

Size of shoe	0	1	**2**	3	4	5	**6**	7	8	9	10
No. of boys	7	10	15	11	18	10	16	5	12	6	2
c f	7	17	32	43	61	71	87	92	104	110	112

Here $N = 112$.

To find Q_1: $m = N/4 = 28$.

The first *c f* just greater than 28 is 32, so the 1st quartile is $Q_1 = 2$

To find Q_3: $m = 3N/4 = 84$.

The first *c f* just greater than 84 is 87, so the 3rd quartile is $Q_3 = 6$

$\therefore$ the semi inter quartile range *i.e.* Q.D. $= \frac{Q_3 - Q_1}{2} = \frac{6-2}{2} = 2$

Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{6-2}{6+2} = 0.5$

Example 7: The following data gives the sales (in '00 Rs.) of two salesmen, Hari and Ganesh in a week. Compare their coefficient of Q.D. of the sales and comment.

Sales of Hari	70	55	80	60	55	88	75
Sales of Ganesh	60	75	85	65	60	90	110

Solution: We first arrange the given data in ascending order.

Observation no.	1	2	3	4	5	6	7
Sales of Hari	55	55	60	70	75	80	88
Sales of Ganesh	60	60	65	75	85	90	110

Here $N = 7$.

For Hari

$(N+1)/4 = 2$

$\therefore Q_1 = 55$

$3(N+1)/4 = 6$

$\therefore Q_3 = 80$

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

$= \frac{80-55}{80+55}$

$= \frac{25}{135}$

$= 0.18$

For Ganesh

$(N+1)/4 = 2$

$\therefore Q_1 = 60$

$3(N+1)/4 = 6$

$\therefore Q_3 = 90$

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

$= \frac{90-60}{90+60} = \frac{30}{150}$

$= 0.2$

Comparing both the coefficients we can say that the deviations in the sales of Hari from the median sales are less than that of the deviations in the sales of Ganesh from his median sales.

6.6.2 Merits and demerits of quartile deviation:

Merits

- It is easy to understand and compute.
- It is rigidly defined.
- It is not affected by extreme observations.
- It can be used for open end class interval distribution.

Demerits

- It is not based on all observations.
- It is affected by sampling fluctuations.
- It cannot be used for further algebraic treatment.

6.7 MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION:

The previous two measures range and Q.D. did not consider all the observations and their deviation from the central value. Mean deviation also called as mean absolute deviation, overcomes this drawback. Mean Deviation (M.D.) is an absolute measure and is defined as the average of all the absolute differences of the observations from the central value. Any of the three averages; mean, median or mode can be taken. Mode of an observation is generally not considered for mean deviation as its value is sometimes indeterminate. The value of M.D. from median is always less than the value of M.D. from mean.

Symbolically the formula for ungrouped and grouped data is as tabled below:

Table 6.1

For ungrouped data	For grouped data
$\frac{\Sigma d}{n}$	$\frac{\Sigma f.d}{N}$
Here $d = \lvert x - \text{avg} \rvert$, n:no. of observations and avg: mean, median or mode	Here $d = \lvert x - \text{avg} \rvert$, $N = \Sigma f$ and avg: mean, median or mode

The relative measure of M.D. is the coefficient of M.D. is given by the formula:

$$\text{Coefficient of M.D.} = \frac{\text{mean deviation}}{\text{average}}$$

The formulae to find the mean deviations are summarized in the table below:

Table 6.2

	Ungrouped Data	Grouped Data	Coefficient of M.D.
M.D. from mean ($\delta\bar{x}$)	$\frac{\Sigma\lvert x-\bar{x}\rvert}{n}$	$\frac{\Sigma f\lvert x-\bar{x}\rvert}{N}$	$\frac{\delta\bar{x}}{\bar{x}}$
M.D. from median (δM)	$\frac{\Sigma\lvert x-M\rvert}{n}$	$\frac{\Sigma f\lvert x-M\rvert}{N}$	$\frac{\delta M}{M}$
M.D. from mode (δZ)	$\frac{\Sigma\lvert x-Z\rvert}{n}$	$\frac{\Sigma f\lvert x-Z\rvert}{N}$	$\frac{\delta Z}{Z}$

Here $\bar{x}$: mean, M: median and Z: mode

Steps to find M.D.

(1) The average from whom the deviation is to be found is computed first.

(2) The absolute differences from the average are calculated and their sum is computed.

(3) In case of grouped data, the product of absolute differences with the corresponding frequencies is calculated and their sum is computed.

(4) Using the appropriate formula the mean deviation is computed.

v *If nothing is mentioned about the average then median is to be taken.*

Example 8: Find the mean deviation from mean and its coefficient for the following data giving the rainfall in cm in different areas in Maharashtra: 105, 90, 102, 67, 71, 52, 80, 30, 70 and 48.

Solution: Since we have to compute M.D. from mean, we first prepare the table for finding mean and then introduce columns of absolute deviations from the mean.

$$\bar{x} = \frac{105+90+102+67+71+52+80+30+70+48.}{10} = \frac{715}{10} = 71.5$$

x	$d = \lvert x-\bar{x}\rvert$
105	33.5
90	18.5
102	30.5
67	4.5
71	0.5
52	19.5
80	8.5
30	41.5
70	1.5
48	23.5
$\sum x = 715$	$\Sigma d = 182$

Now,

$$\text{M.D. from mean} = \frac{\Sigma d}{n}$$

$$\delta\bar{x} = \frac{182}{10}$$

$$\therefore \delta\bar{x} = 18.2$$

Coefficient of M.D. from mean

$$= \frac{\delta\bar{x}}{\bar{x}} = \frac{18.2}{71.5}$$

Example 9: The marks obtained by 10 students in a test are given below. Find the M.D. from median and its relative measure.

Marks: 15 10 10 03 06 04 11 17 13 05

Solution: The marks of 10 students are arranged in ascending order and its median is found. The column of absolute deviations from median is introduced and its sum is computed. Using the formula mentioned above, M.D. from median and its coefficient is calculated.

x	$d = \lvert x - M \rvert$
03	5
04	4
05	3
06	2
10	2
10	2
11	3
13	5
15	7
17	9
Total	$\Sigma d = 42$

Since $N = 10$,
Median = A.M. of 5th and 6th observation

$$\therefore M = \frac{6+10}{2} = 8$$

M.D. from median $= \frac{\Sigma d}{n} = \frac{42}{10}$

$\therefore \delta M = 4.2$

Coefficient of M.D. $= \frac{\delta M}{M} = \frac{4.2}{8} = 0.525$

Example 10: On the Mumbai – Nashik highway the number of accidents per day in 6 months are given below. Find the mean deviation and coefficient of M.D.

No. of accidents	0	1	2	3	4	5	6	7	8	9	10
No. of days	26	32	41	12	22	10	05	01	06	15	10

Solution:

No. of accidents	No. of days (f)	cf	$d = \lvert x - M \rvert$	$f.d$
0	26	26	2	52
1	32	58	1	32
2	41	99	0	0
3	12	111	1	12
4	22	133	2	44

5	10	143	3	30
6	05	148	4	20
7	01	149	5	05
8	06	155	6	36
9	15	170	7	105
10	10	180	8	80
Total	$N = 180$	-	-	$\Sigma fd =$ 416

Now, $N = 180 \therefore m = N/2 = 180/2 = 90$.

The cf just greater than 90 is 99. The corresponding observation is 2. $\therefore M = 2$

M.D. from median = $\delta M = \frac{\Sigma fd}{N} = \frac{416}{180} = 2.31$

Coefficient of M.D. = $\frac{\delta M}{M} = \frac{2.31}{2} = 1.15$

Example 11: The following data gives the wages of 200 workers in a factory with minimum wages Rs. 60 and maximum wages as Rs. 200. Find the mean deviation and compute its relative measure.

Wages less than	80	100	120	140	160	180	200
No of workers	30	45	77	98	128	172	200

Solution: The data is given with less than *cf*, we first convert them to frequencies then find the median and follow the steps to compute M.D. as mentioned above.

Wages in Rs	No. of workers (f)	cf	x	$d = \lvert x - 141.33 \rvert$	fd
60 – 80	30	30	70	71.33	2139.9
80 – 100	15	45	90	51.33	769.95
100 – 120	32	77	110	31.33	1002.56
120 – 140	21	**98**	130	11.33	237.93
140 – 160	**30**	128	150	8.67	260.1
160 – 180	44	172	170	28.67	1261.48
180 – 200	28	200	190	48.67	1362.76
Total	$N = 200$	-	-	-	$\Sigma fd = 7034.68$

Now, $N = 200 \therefore m$ and $= N/2 = 200/2 = 100$.

The median class is 140 – 160. $\therefore l_1 = 140$, $l_2 = 160$, $i = 160\text{-}140 = 20$, $f = 30$ and $pcf = 98$

$$\therefore M = l_1 + \left[\frac{m - pcf}{f} \text{ x } i\right] = 140 + \left[\frac{100 - 98}{30} \text{ x } 20\right] = 140 + 1.33 = 141.33$$

M.D. from median $= \delta M = \frac{\Sigma fd}{N} = \frac{7034.68}{200} = 35.1734$

Coefficient of M.D. $= \frac{\delta M}{M} = \frac{35.1734}{141.33} = 0.25$

Example 12: The following data gives the ages of people residing in a society. Find the mean, M.D. from mean and coefficient of M.D.

Ages in yrs.	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No of people	21	16	43	54	32	12	22

Solution: The table for computing mean and M.D. from mean is as follows:

Ages in yrs.	Mid point (x)	No. of people (f)	Fx	d	fd
0 – 10	5	21	105	29.2	613.2
10 – 20	15	16	240	19.2	307.2
20 – 30	25	43	1075	9.2	395.6
30 – 40	35	54	1890	0.8	43.2
40 – 50	45	32	1440	10.8	345.6
50 – 60	55	12	660	20.8	249.6
60 – 70	65	22	1430	30.8	677.6
Total	-	200	$\Sigma fx =$ 6840	-	$\Sigma fd =$ 2632

From the table we have $N = 200$ and $\Sigma fx = 6840$

$$\therefore \bar{x} = \frac{\Sigma fx}{N} = \frac{6840}{200} = 34.2$$

We compute the absolute differences of the mid points from $\bar{x}$ in the 5th column of d and multiply with the corresponding frequencies in the 6th column of fd.

Now, we have $\Sigma fd = 2632$ and $N = 200$

$\therefore$ M.D. from mean = $\delta\bar{x} = \frac{\Sigma fd}{N} = \frac{2632}{200} = 13.16$

Coefficient of M.D. from mean = $\frac{\delta\bar{x}}{\bar{x}} = \frac{13.16}{34.2} = 0.38$

6.7.1 Merits and demerits of mean deviation:

Merits

- It is easy to understand and compute.
- It is rigidly defined.
- It is based on all observations.
- It is not affected by extreme observations.

Demerits

- Though it is rigidly defined, it can be calculated using any of the averages, which may create problems in comparing different mean deviations.
- If mean deviation is calculated from mode it does not prove to be an accurate measure of dispersion.
- It cannot be used for further algebraic treatment.

6.8 VARIANCE AND STANDARD DEVIATION:

In computing mean deviations the sign of the deviations are ignored by taking their absolute value. This can be overcome by taking the squares of the deviations.

The average of the square of the deviations measured from mean is called as *variance.*

Symbolically, variance = $\frac{\Sigma d^2}{n}$ or $\frac{\Sigma f.d^2}{N}$; where $d = x - \bar{x}$.

The positive square root of variance is called as standard deviation (S.D.) and is denoted by the Greek alphabet 'σ' (sigma).

Symbolically, $\sigma = \sqrt{\frac{\Sigma d^2}{n}}$ or $\sigma = \sqrt{\frac{\Sigma f.d^2}{N}}$; where $d = x - \bar{x}$.

…(2.1)

Thus, σ^2 = variance.

S.D. can also be computed by another formula:

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} \qquad \ldots(2.2)$$

$$\sigma = \sqrt{\frac{\Sigma f.x^2}{N} - (\bar{x})^2} = \sqrt{\frac{\Sigma f.x^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2} \qquad \ldots(2.3)$$

The advantage of second type of formulae over (1) is that, if the number of observations is large and the mean is not a whole number, it is easier to find x^2 and $f.x^2$ than calculating the absolute deviations (d) from mean and their squares.

Steps to find S.D. *using formula* (2):

Ungrouped Data	Grouped Data
1. Find the sum of all observations. *i.e.* Σx.	1. Find the class mid points; multiply by the corresponding frequencies and total it. *i.e.* find Σfx.
2. Square the observations and find its total, *i.e.* Σx^2.	2. Multiply each entry from the column of fx with x to get fx^2 and sum them to get $\Sigma f.x^2$.
3. Using the formula (2.2), S.D. is computed.	3. Using the formula (2.3), S.D. is computed.

v In the following solved examples only the first example is solved using formula **2.1**, rest all are solved using formula **2.2** or **2.3**. Students can themselves understand the simplicity and speed of these formulae over **2.1**.

Example 13: The marks of internal assessment obtained by FYBMS students in a college are given below. Find the mean marks and standard deviation.

22 30 36 12 15 25 18 10 33 29

Solution: We first sum all the observations and find the mean. Then the differences of the observations from the mean are computed and squared. The positive square root average of sum of square of the differences is the required standard deviation.

x	$d = x - \bar{x}$	d^2
22	-1	1
30	7	49

I. $\bar{x} = \frac{\Sigma x}{n} = \frac{230}{10} = 23$.

II. $\sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{738}{10}}$

36	13	169
12	-11	121
15	-8	64
25	2	4
18	-5	25
10	-13	169
33	10	100
29	6	36
$\Sigma x = 230$	-	$\Sigma d^2 = 738$

Example 14: Find the standard deviation for the following data:

03 12 17 29 10 05 18 14 12 20

Solution: We find the sum of the observations and the sum of its squares. Using formula **2.2**, S.D. is computed as follows:

x	x^2
3	9
12	144
17	289
29	841
10	100
05	25
18	324
14	196
12	144
20	400
$\Sigma x = 140$	$\Sigma x^2 = 2472$

The formula **2.2** gives the S.D. as below:

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{2472}{10} - \left(\frac{140}{10}\right)^2}$$

$$= \sqrt{247.2 - 196} = \sqrt{51.2}$$

Example 15: Compute the standard deviation for the following data

x	100	102	104	106	108	110	112
F	5	11	7	9	13	10	12

Solution: *Short-cut method*:

In problems where the value of x is large (consequently its square also will be very large to compute), we use the short-cut method. In this method, a

fixed number x_0 (which is usually the central value among x) is subtracted from each observation. This difference is denoted as

$u = x - x_0$. Now the columns of fu and fu^2 are computed and the S.D. is calculated by the formula: $\sigma = \sqrt{\frac{\Sigma f.u^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2}$. One can observe that this formula is similar to that mentioned in **2.3**. This formula is called as change of Origin formula.

In this problem we assume $x_0 = 106$. The table of calculations is as follows:

x	$u = x - 106$	f	fu	fu^2
100	-6	5	-30	180
102	-4	11	-44	176
104	-2	7	-14	28
106	0	9	0	0
108	2	13	26	52
110	4	10	40	160
112	6	12	72	432
Total	-	$N = 67$	$\Sigma fu = 50$	$\Sigma f.u^2 = 1028$

In the table, the column of fu^2 is computed by multiplying the entries of the columns fu and u.

From the table we have: $\Sigma f.u^2 = 1028$, $\Sigma fu = 50$ and $N = 67$.

$$\therefore \sigma = \sqrt{\frac{\Sigma f.u^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} = \sqrt{\frac{1028}{67} - \left(\frac{50}{67}\right)^2}$$

$$\therefore \sigma = \sqrt{15.34 - 0.56} = \sqrt{14.78}$$

$$\therefore \sigma = 3.84$$

Example 16: The income (in '000 Rs.) of 100 families is given below. Find the mean and S.D.

Income	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
No. of families	18	20	26	5	10	12	9

Solution: Computation using short-cut method:

Income	x	$u = x - 17.5$	f	fu	fu^2
0 – 5	2.5	-15	18	-270	4050

5 – 10	7.5	-10	20	-200	2000
10 – 15	12.5	-5	26	-130	650
15 – 20	17.5	0	5	0	0
20 – 25	22.5	5	10	50	250
25 – 30	27.5	10	12	120	1200
30 – 35	32.5	15	9	135	2025
Total	-	-	$N =$ 100	$\Sigma fu =$ - 295	$\Sigma f.u^2 =$ 10175

From the table we have: $\Sigma f.u^2 = 10175$, $\Sigma fu = -295$ and $N = 100$.

$$\therefore \sigma = \sqrt{\frac{\Sigma f.u^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} = \sqrt{\frac{10175}{100} - \left(\frac{-295}{100}\right)^2}$$

$$= \sqrt{101.75 - 8.7025} = \sqrt{93.0475}$$

$\therefore \sigma =$ Rs. 9.65

v Sometimes only change of origin does not simplify the calculations. In such cases change of scale method is applied. In this method, we assume $u = \frac{x - x_0}{i}$, where i : change of scale. The formula for mean and S.D. using this u is not the same. $\bar{x} = x_0 + i.\bar{u}$ and $\sigma_x = i$ x σ_u, where $\sigma_u = \sqrt{\frac{\Sigma f.u^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2}$

Example 17: The following data gives scholarships awarded to students of a college. Find the S.D.

Scholarship	1000	2000	3000	4000	5000
No. of students	16	20	8	10	6

Solution: Here we take $x_0 = 3000$ and $i = 1000$

Scholarship	$u = \frac{x - 3000}{1000}$	f	fu	fu^2
1000	-2	16	-32	64
2000	-1	20	-20	20
3000	0	8	0	0
4000	1	10	10	10

5000	2	6	12	24
Total	-	N = 60	Σfu = -30	Σfu^2 = 118

From the table we have, $N = 60$, $\Sigma fu = -30$ and $\Sigma fu^2 = 118$.

Now, $\sigma_u = \sqrt{\frac{\Sigma f.u^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} = \sqrt{\frac{118}{60} - \left(\frac{30}{60}\right)^2} = \sqrt{1.96 - 0.25} = \sqrt{1.71} = 1.3076$

$\sigma_x = i$ x σ_u = 1000 x 1.3076 = 1307.66

$\therefore \sigma_x$ = Rs. 1308 (approx.)

6.8.1 Coefficient of variation:

Standard Deviation is an absolute measure of dispersion and is expressed in the same units of measurement as that of the data. To compare two different sets of data we need a relative measure which is free from the units of the observations. The relative measure of dispersion for standard deviation given by Karl Pearson, is called as *coefficient of standard deviation* or also as coefficient of variation (*CV*). The formula to compute *CV* is as follows:

Coefficient of variation (*CV*) = $\frac{\sigma}{\bar{x}}$ x 100

This relative measure of S.D. measure how large is the S.D. in comparison with the mean of the data. The data whose *CV* is small is said to be more consistent.

Example 18: Find the mean, S.D. and *CV* for the following data:

Solution:

C.I.	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
f	3	9	7	2	4

C.I.	x	$u = \frac{x - 22.5}{5}$	f	fu	fu^2
10 – 15	12.5	-2	3	-6	12
15 – 20	17.5	-1	9	-9	9
20 – 25	22.5	0	7	0	0
25 – 30	27.5	1	2	2	2
30 – 35	32.5	2	4	8	16
Total	-	-	N = 25	Σfu = -5	Σfu^2 = 39

From the table: $N = 25$, $\Sigma fu = -5$, $\Sigma fu^2 = 39$.

$$\bar{x} = x_0 + i.\bar{u} = 22.5 + 5\left(\frac{-5}{25}\right) = 22.5 - 1 = 21.5 \qquad \therefore \bar{x} = 21.5 \qquad \ldots (1)$$

$$\sigma_u = \sqrt{\frac{\Sigma f.u^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} = \sqrt{\frac{39}{25} - \left(\frac{-5}{25}\right)^2} = \sqrt{1.56 - 0.04} = \sqrt{1.52} = 1.2328$$

$$\therefore \sigma_x = i \text{ x } \sigma_u = 5(1.2328) = 6.164 \qquad \therefore \sigma_x = 6.164 \qquad \ldots (2)$$

Coefficient of Variation $CV = \frac{\sigma}{\bar{x}}$ x $100 = \frac{6.164}{21.5}$ x $100 = 28.66\,\%$ … (3)

Example 19: Choudhari & Bros own a factory which manufactures two products *A* and *B*. The profit (in '000 Rs.) of the two products from 1995 to 2003 is given below. Find which product gives more consistent profit to Choudhari & Bros.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003
Profit for *A*	101	95	110	105	112	99	102	100	112
Profit for *B*	82	90	109	80	81	72	75	80	115

Solution:

Profit for Product *A*			Profit for Product *B*		
x	$u = x - 102$	u^2	x	$u = x - 80$	u^2
101	-1	1	82	2	4
95	-7	49	90	10	100
110	8	64	109	19	361
105	3	9	80	0	0
112	10	100	81	1	1
99	-3	9	72	-8	64
102	0	0	75	-5	25
100	-2	4	80	0	0
112	10	100	115	35	1225
Total	$\Sigma u = 18$	$\Sigma u^2 = 336$	Total	$\Sigma u = 54$	$\Sigma u^2 = 1780$

For Product *A*:

$\bar{x} = x_0 + \bar{u}$ = 102 + 18/9 = 102 + 2 = 104

$$\sigma = \sqrt{\frac{\Sigma u^2}{N} - \left(\frac{\Sigma u}{N}\right)^2} = \sqrt{\frac{336}{9} - \left(\frac{18}{9}\right)^2} = \sqrt{37.33 - 4} = \sqrt{33.33} = 5.7735$$

Now, $CV = \frac{\sigma}{\bar{x}}$ x 100 = $\frac{5.7735}{104}$ x 100 = 5.55 … (1)

For Product *B*:

$\bar{x} = x_0 + \bar{u}$ = 102 + 54/9 = 102 + 6 = 108

$$\sigma = \sqrt{\frac{\Sigma u^2}{N} - \left(\frac{\Sigma u}{N}\right)^2} = \sqrt{\frac{1780}{9} - \left(\frac{54}{9}\right)^2} = \sqrt{197.77 - 36} = \sqrt{161.77} = 12.7192$$

Now, $CV = \frac{\sigma}{\bar{x}}$ x 100 = $\frac{12.7192}{108}$ x 100 = 11.78 … (1)

From (1) and (2), we observe that the coefficient of variation of product *A* is less than that of product *B*.

Thus, the profits earned by product *A* are more consistent.

Example 20: The average pulse rate of patient increased from 62 to 70 and the S.D. also increased from 0.8 to 1.2 after treatment. Is it right to conclude that there is improvement in the patients' health?

Solution: Given $\bar{x}_1$ = 62, σ_1 = 0.8 and $\bar{x}_2$ = 70, σ_2 = 1.2

$$\therefore CV_1 = \frac{\sigma_1}{\bar{x}_1} \text{ x } 100 = \frac{0.8}{62} \text{ x } 100 = 1.29$$

After treatment, $\therefore CV_2 = \frac{\sigma_2}{\bar{x}_2}$ x 100 = $\frac{1.2}{72}$ x 100=1.66

The *CV* after treatment is higher than before treatment. This means after treatment the pulse rate have become more variable. Hence it is not proper to conclude that the health of the patient has improved.

6.8.2 Combined standard deviation:

One of the properties of standard deviation is that it can be used for further algebraic treatment. We can find the combined standard deviation of two different sets of data.

Consider two sets of data with number of observations n_1, n_2; their respective means x_1, x_2 and respective standard deviations as σ_1 and σ_2. Then the combined standard deviation is given by the formula:

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where $d_1 = \bar{x}_1 - \bar{x}_{12}$, $d_2 = \bar{x}_2 - \bar{x}_{12}$ and $\bar{x}_{12} = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$ the combined mean.

Example 21: Find the combined mean and S.D. for the following two groups with the details given below. Also find which group is more variable.

	Group I	Group II
observations	40	60
Mean	60	70
S.D.	8	5

Solution: Given $n_1 = 40, \bar{x}_1 = 60, \sigma_1 = 8$ and $n_2 = 60, \bar{x}_2 = 70$ and $\sigma_2 = 5$

Combined mean = $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{40(60) + 60(70)}{40 + 60} = \frac{6600}{100} = 66$

Now, $d_1 = \bar{x}_1 - \bar{x}_{12}$ = 60 – 66 = -6 and $d_2 = \bar{x}_2 - \bar{x}_{12}$ = 70 – 66 = 4

Combined S.D. = $\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}} =$

$\sqrt{\frac{40(8^2 + (-6)^2) + 60(5^2 + 4^2)}{40 + 60}}$

$\therefore \sigma_{12} = \sqrt{\frac{40(64 + 36) + 60(25 + 16)}{100}} = \sqrt{\frac{4000 + 2460}{100}} = \sqrt{64.6} = 8.04$ (approx.)

To find which group is variable, we compute the *CV* of the both the groups.

For Group I: $CV_1 = \frac{\sigma_1}{\bar{x}_1} \text{ x } 100 = \frac{800}{60} = 13.33$

For Group II: $CV_2 = \frac{\sigma_2}{\bar{x}_2} \text{ x } 100 = \frac{500}{70} = 7.14$

The *CV* of the Group I is higher than that of Group II. Thus, Group I is more variable than Group II.

Example 22: If the coefficient of variation of two groups is 13.25% and 17% and their averages are 110 and 95 respectively, find the corresponding standard deviations.

Solution: Given $CV_1 = 13.25$, $\bar{x}_1 = 110$ and $CV_2 = 17$, $\bar{x}_2 = 95$

If σ_1 and σ_2 are the corresponding standard deviations, using the formula for *CV*, we have

$CV_1 = \frac{\sigma_1}{\bar{x}_1} \text{ x } 100 \qquad \Rightarrow \sigma_1 = \frac{\bar{x}_1 . CV_1}{100} = \frac{110(13.25)}{100} = 14.58$

$CV_2 = \frac{\sigma_2}{\bar{x}_2}$ x 100 $\Rightarrow \sigma_2 = \frac{\bar{x}_2 . CV_2}{100} = \frac{95(17)}{100} = 16.15$

CORRECTED S.D.: Just as the arithmetic mean can be corrected for the incorrect observations, S.D. also can be corrected. Before understanding the steps to find the correct S.D. let us understand the formula of S.D. in more detail.

We know that, $\sigma = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}$ $\Rightarrow \sigma^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2$

$\therefore \Sigma x^2 = n(\sigma^2 + \bar{x}^2)$.

Thus, if any of the observations is incorrect the value of Σx^2 becomes incorrect and has to be corrected first. Now we shall write down the steps to calculate the correct S.D.

Steps to find correct S.D.

(1) We calculate the wrong Σx value by using the formula: $\Sigma x = n\bar{x}$.

(2) Now we calculate the wrong Σx^2 by using the formula: $\Sigma x^2 = n(\sigma^2 + \bar{x}^2)$

(3) Correct Σx = (Σx) – wrong observation + correct observation.

(4) Correct Σx^2 = (Σx^2) – (wrong observation)^2 + (correct observation)^2.

(5) Now the correct mean and standard deviation is computed using the formulae:

Correct $\bar{x} = \frac{\text{correct } \Sigma x}{n}$ and Correct $\sigma = \sqrt{\frac{\text{correct } \Sigma x^2}{n} - (\text{correct } \bar{x})^2}$

This method can be similarly extended to problems where there are more than one wrong observations.

Example 23: For a certain data with 10 observations, the mean and S.D. were calculated to be 9 and 4 respectively. But later on it was observed that one of the value was wrongly taken as 14 instead of 4. Find the correct mean and the correct S.D.

Solution: Given: n = 10, $\bar{x}$ = 9, σ = 4, wrong value = 14 and correct value = 4

$\Sigma x = n\bar{x}$ = 10(9) = 90

$\Sigma x^2 = n(\sigma^2 + \bar{x}^2)$ = 10(16 + 81) = 970

Now, correct $\Sigma x = 90 - 14 + 4 = 80$

and correct Σx^2 = 970 – $(14)^2 + 4^2$ = 970 – 196 + 16 = 790

Correct $\bar{x} = \frac{\text{correct } \Sigma x}{n} = 80/10 = 8$

Correct S.D. $= \sigma = \sqrt{\frac{\text{correct } \Sigma x^2}{n} - (\text{correct } \bar{x})^2} =$

$\sqrt{\frac{790}{10} - (8)^2} = \sqrt{79 - 64} = \sqrt{15} = 3.87$

Example 24: A problem of finding the mean and standard deviation was given to the students of a class by their teacher. The mean and S.D. of 20 observations was calculated as 12 and 4 respectively. Later on the teacher found that one of the observations was misheard by students as 13 instead of 30. Find the correct mean and S.D.

Solution: Given: $n = 20$, $\bar{x} = 12$, $\sigma = 4$, wrong value = 13 and correct value = 30

$\Sigma x = n\bar{x} = 20(12) = 240$

$\Sigma x^2 = n(\sigma^2 + \bar{x}^2) = 20(16 + 144) = 3200$

Now, correct $\Sigma x = 240 - 13 + 30 = 257$

and correct $\Sigma x^2 = 3200 - (13)^2 + (30)^2 = 3200 - 169 + 900 = 3931$

Correct $\bar{x} = \frac{\text{correct } \Sigma x}{n} = 257/20 = 12.85$

Correct S.D. $= \sigma = \sqrt{\frac{\text{correct } \Sigma x^2}{n} - (\text{correct } \bar{x})^2}$

$= \sqrt{\frac{3931}{20} - (12.85)^2} = \sqrt{196.55 - 165.1225} = \sqrt{31.4275}$

$\therefore$ correct $\sigma = 5.61$

6.8.3 Merits and demerits of standard deviation:

Merits

(1) It is based on all observations.

(2) It is rigidly defined.

(3) It can be used for further algebraic treatment.

(4) It is not affected by sampling fluctuations.

Demerits

(1) Compared with other measures of dispersion, it is difficult to calculate.

(2) More importance is given to the extreme observations in calculating standard deviation. The square of the deviations of extreme values from the mean, dominate the total and hence the value of S.D.

MISSING VALUES

Example 25: Find the missing values in the following table:

	Group I	Group II	Combined
observations	60	?	140
mean	?	22.5	30
S.D.	4.5	7	?

Solution: Given: $n_1 = 60$ and $n_1 + n_2 = 140 \Rightarrow n_2 = 140 - 60 = 80$.

Also given: $x_{12} = 30$, $x_2 = 22.5$, using the combined mean formula, we have

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \qquad \Rightarrow 30 = \frac{60\bar{x}_1 + 80(22.5)}{140}$$

$$\Rightarrow 4200 = 60\bar{x}_1 + 1800 \qquad \therefore 60\bar{x}_1 = 2400$$

$$\therefore \bar{x}_1 = 40$$

Now, $d_1 = \bar{x}_1 - \bar{x}_{12} = 40 - 30 = 10$ and $d_2 = \bar{x}_2 - \bar{x}_{12} = 22.5 - 30 = -7.5$

Combined S.D. $\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}} =$

$$\sqrt{\frac{60[(4.5)^2 + 10^2] + 80[7^2 + (-7.5)^2]}{60 + 80}}$$

$$\sigma_{12} = \sqrt{\frac{7215 + 8420}{140}} = 10.56$$

CHOICE OF A MEASURE OF DISPERSION

Throughout the chapter we have seen different measures of dispersion like range, inter quartile range, quartile deviation, mean deviation and standard deviation with their merits and demerits. The selection of a particular measure of dispersion depends upon the following three aspects:

(1) *The objective to find the measure of dispersion*: If the purpose is to find the degree of variation of the observations from the mean then standard deviation is more suitable than mean deviation.

(2) *The type of data available*: If the data available has open end class intervals then it is not possible to use mean deviation or standard deviation. Also if the data is very large and scattered, range cannot be used as a proper measure of dispersion.

(3) *The characteristics of the measure of dispersion*: The merits and demerits of the different measure over one another would be helpful to select the most required one.

6.9 SKEWNESS& KURTOSIS:

Definition: Skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve which can be drawn with the help of given data.

Distribution is said to be skewed if –

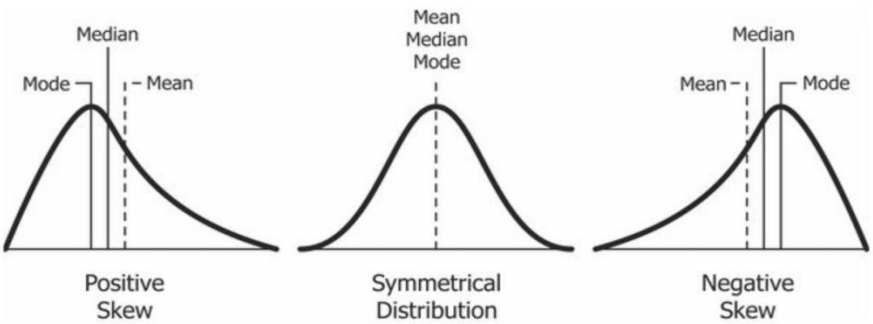
1. Mean, median and mode fall at different points.
2. Quartiles are not equidistance from median; and
3. The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to other.

Note:

1. A distribution is said to be **symmetric** about its arithmetic mean (A.M.) if the deviation of the values of the distribution from their A.M. are such that corresponding to each positive deviation, there is negative deviation of the same magnitude.
2. If the distribution is **symmetric** then $\mu_3 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{n} = 0$ & $\mu_3 = \frac{\sum_{i=1}^{n} f_i(x_i-\bar{x})^3}{N} = 0$.
3. If a distribution is not symmetric then the distribution is called a skewed distribution.
4. A skewed distribution is also called as asymmetric distribution.
5. Thus in case of a **skewed distribution** the magnitudes of the positive and the negative deviations of the values from their mean do not balance.

Types of Skewness:

Skewness is of two types



(Fig. 01)

1. **Positive Skewness**: Skewness is **positive** if the larger tail of the distribution lies towards the higher values of the variate (the right), i.e. if the curve drawn with the help of the given data is stretched more to the right than left and the distribution is said to be **positively skewed distribution**.

   For a positively asymmetric distribution: **A.M. > Median > Mode**

2. **Negative Skewness**: Skewness is **negative** if the larger tail of the distribution lies towards the lower values of the variate (the left), i.e. if the curve drawn with the help of the given data is stretched more to the left than right and the distribution is said to be **negatively skewed distribution**.

   For a negatively asymmetric distribution: **A.M. < Median < Mode**

Note:

1. For a symmetric distribution: A.M. = Median = Mode
2. Skewness is positive if A.M. > Median or A.M. > Mode
3. Skewness is negative if A.M. < Median or A.M. < Mode

Measure of Skewness

Various measures of skewness are (these are absolute measures of skewness)

1. S_k = Mean – Median
2. S_k = Mean – Mode
3. Sk = $(Q_3 – M_d) – (M_d – Q_1)$

Kurtosis

The three measures namely, measures of central tendency, measure of variations (moments) and measure of skewness that we have studied so far are not sufficient to describe completely the characteristics of a frequency distribution. Neither of these measures is concerned with the peakedness of a frequency distribution.

Kurtosis is concerned with the flatness or peakedness of frequency curve – The graphical representation of frequency distribution.

Definition:

Clark and Schkade defined kurtosis as: "Kurtosis is the property of a distribution which express its relative peakedness."

Types of Kurtosis:

1. **Mesokurtic**
2. **Leptokurtic**

3. **Platykurtic**

1. **Mesokurtic:** The frequency curve which is bell shaped curve is considered as standard and such distribution is called Mesokurtic.

   The normal curve is termed Mesokurtic.

2. **Leptokurtic:** A curve which is more peaked than the normal curve is called Leptokurtic. For **Leptokurtic curve kurtosis is positive** and dispersion is least among all the three types.

3. **Platykurtic:** A curve which is flatter than the normal curve is called Platykurtic. For **Platykurtic curve kurtosis is negative** and dispersion is more.

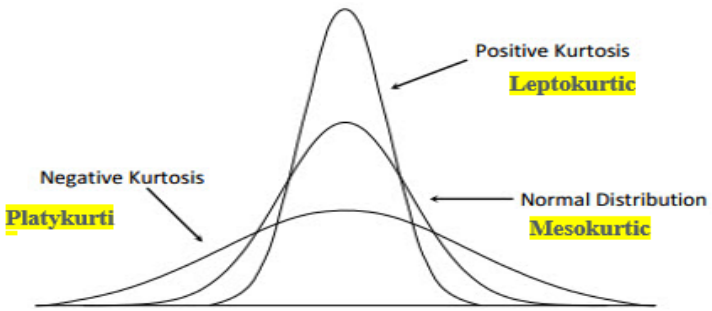
6.10 LET US SUM UP:

In this chapter we have learn:

- How to calculate Range and coefficient of range.
- How to calculate Quartile deviation and coefficient of quartile deviation.
- How to calculate Mean deviation and coefficient of Mean deviation.
- How to calculate Standard deviation and coefficient of variation.

6.11 UNIT END EXERCISES:

1. Define Dispersion. Discuss the importance of different measures of Dispersion.
2. What are the functions of a measure of dispersion?
3. Explain with suitable example how does a measure of dispersion



   proves to be a supplementary tool for averages.
4. Differentiate between relative and absolute dispersion.

5. Define Mean Deviation. Illustrate with examples different types of mean deviations.

6. Write a short note on variance and its advantages over other measures of dispersion.

7. Discuss with suitable examples the advantages of coefficient of variation.

8. What are the criteria to select a particular measure of dispersion?

9. Explain briefly the different measures of dispersion.

10. Define Quartile Deviation. Explain the advantages of Q.D. over range.

11. Write a short note on standard deviation and explain with examples why is it the most popular measure of dispersion.

12. Define coefficient of variation. Explain its importance in statistical analysis of series of data.

13. Find the range and its coefficient for the following data:

    a) Wages in Rs.: 100, 55, 45, 90, 80, 120, 30, 125, 140 and 40.

    b) Marks: 78, 37, 56, 89, 22, 30, 34, 10, 55, 38, 46, 62, 77, 12 and 44.

    c) Temperature in degree Celsius: 32.5, 33.8, 32, 35, 35.2, 38, 33, 32.7, 31 and 31.4

    d) Share Price in Rs: 1021, 1000, 1009, 1022, 1022.5, 1024, 1011, 1015, 1002.5, 1020

    e) Rainfall in mm: 65, 67, 77, 62, 60, 56, 60, 45, 76, 80 and 44.

14. Find the range and its coefficient for the following data:

Income	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
No. of families	12	14	11	18	20	08	10

15. If the coefficient of range is 0.5 and the smallest value is 10, find the largest value of the data.

16. If the coefficient of range is 0.8 and the largest value is 40, find the smallest value of the data.

17. Find the inter quartile range for the following data:

    a) Marks: 5, 10, 7, 2, 8, 11, 16, 1, 6, 11, 12, 3, 7, 14, 12, 10

b)

Fuel in ltrs	1 – 5	6 – 10	11 – 15	16 – 20	21 – 25	26 – 30
No. of vehicles	40	35	10	15	20	30

18. Find the Q.D. for the following data and also find its coefficient.

Height in cm	120	125	130	135	140	150	160
No. of boys	6	12	18	22	32	10	5

19. Find the Q.D. for the following data and also find its coefficient.

No. of phone calls	100	120	140	160	180	200	220
Frequency	5	15	18	22	19	10	6

20. Find the Q.D. for the following data and also find its coefficient.

Income in Rs. less than	50	70	90	110	130	150	170
No of families	54	150	290	595	820	900	1000

21. Find the range of the central 50 % of workers, Q.D. and its coefficient for the following data:

Income in Rs.	10 – 30	30 – 50	50 – 70	70 – 90	90 – 110	110-130	130-150
No of workers	7	18	12	15	10	8	10

22. Find the M.D. from median for the following data. Also compute its coefficient.

Size	4	8	12	16	20	24	28	32
Frequency	6	11	17	20	12	14	8	10

23. The following data gives the rainfall in cm in past 10 years for two cities *A* and *B*. Find the M.D. from mean and coefficient of M.D. and comment.

City *A*	36	67	35	50	54	71	60	45

City *B*	80	85	77	79	80	82	88	85

24. Compute the M.D. from mean, median and mode for the following data. Compare the three values, which M.D. is the lowest?

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
No of students	16	20	15	24	10	38	22	15

25. Compute the mean deviation from mean and median for the following data. Also find the coefficient of mean deviation.

C.I.	3 – 4	4 – 5	5 – 6	6 – 7	7 – 8	8 – 9	9 – 10
Frequency	10	16	14	19	12	4	2

26. The differences between the ages of husband and wives are given below. Find the M.D. from mean and also find the coefficient of M.D.

Age in yrs	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12
No of couples	349	278	402	100	112	25

27. Compute the standard deviation for the following data giving marks obtained by FYBMS student in their Statistics project.

12	8	10	11	7	3	19	12	12	13	7	5	10	6	4
8	11	15	5	16	7	13	14	11	5	9	6	4	2	10

28. Compute the standard deviation for the following data:

Age in yrs	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30
No of people	26	11	05	08	28	30

29. Calculate the mean and standard deviation for the following data:

C.I.	10 – 12	12 – 14	14 – 16	16 – 18	18 – 20	20 – 22	22 – 24
Frequency	100	102	89	132	90	88	55

30. Calculate the mean and standard deviation for the following data:

Marks less than	15	30	45	60	75	90
No of students	40	60	50	10	30	10

31. Calculate the mean and standard deviation for the following data:

length of wire in cm	12 – 15	16 – 19	20 – 23	24 – 27	28 – 31	32 – 35
Frequency	16	21	18	20	14	11

32. The speed of vehicles on the Mumbai – Pune express Highway per day is given below.

Find the mean and S.D.

Speed in km	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90-100
No of vehicles	78	105	150	177	188	200	170

33. After a month long control of diet and exercises the loss in weight of 50 people in a society is given below. Find the mean and S.D. for the data

Weight loss in kg	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12
No of persons	5	10	12	10	5	8

34. The amount of money taken from an ATM machine is as given below. Find the S.D.

Amount in '000 Rs.	2 – 5	5 – 8	8 – 11	11 – 14	14 – 17	17 – 20
No of days	15	37	25	28	40	30

35. Compute the mean and S.D. for the following data:

Age below	10	20	30	40	50	60	70
f	20	25	40	15	30	25	10

36. The following data gives the number of goals made by the Indian and Pakistani Hockey team. Find out which team is more consistent.

Goals	0	1	2	3	4	5
Indian Team	20	18	10	7	3	5
Pakistan Team	25	15	16	4	3	2

37. The following data gives the duration of phone calls made by boys and girls of a college. Find the S.D., CV.

Duration in seconds	0 – 60	60 – 120	120 – 180	180 – 240	240-300	300-360
No of Boys	17	20	30	16	10	15
No of Girls	20	30	24	18	8	10

38. The life of two types of tube lights in market is given below. Which type of tube light is more uniform?

Life in months	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24
Tube A	8	12	18	6	10	4
Tube B	12	20	32	24	16	10

39. The average weekly wages of workers in two factories are Rs. 85 and Rs. 60 respectively. The number of workers is 100 and 120 while the standard deviation of wages is Rs. 12 and Rs. 8 respectively. Find the combined mean and standard deviation.

40. The mean and standard deviations of two groups with 100 and 150 observations are 55,6 and 40,8. Find the combined mean and S.D. for all the observations taken together.

41. The average weekly wages of 60 workers in a firm A are Rs. 210 with S.D. Rs. 10, while in firm B the numbers are 100, Rs. 90 and Rs. 12. Which firm has greater consistency in the weekly wages? Find the combined wages and S.D. for all the workers.

42. The mean and standard deviations of two groups with 50 and 70 observations are 110,9 and 80,8. Find the combined mean and S.D. for all the observations taken together.

43. The average performance and S.D. of two machines in a factory are 80, 5 and 75, 5.5 respectively. Which machine is more consistent?

44. A sample of size 20 has mean 4.5 and S.D. 2.8. Another sample of size 25 has mean 5.6 and S.D. 3. Find the mean and S.D. of the combined samples.

45. Find the missing values in the following table:

	Group I	Group II	Combined
observations	25	?	60
mean	?	13.5	12
S.D.	?	8	10

46. Two samples of size 30 and 70 have same mean 40 but different S.D. 14 and 18 respectively. Find the combined S.D. of the sample of total size.

47. Find the missing values in the following table:

	Group I	Group II	Combined
observations	?	30	50
mean	28	?	32
S.D.	4.5	?	6

48. For a distribution of 300 observations the mean and S.D. was found to be 50 and 5 respectively. Later on it was found that one of the observations was wrongly taken as 15 instead of 50. Find the correct mean and S.D.

49. The mean and S.D. of 100 observations was calculated to be 35 and 3.5 respectively. One of the observations was wrongly taken as 14. Calculate the mean and S.D. if (*i*) the wrong value is omitted and (*ii*) it is replaced by the correct value 40.

50. A group of 50 observations has mean 65cm and S.D. 8. Two more observations 70

Multiple choice questions:

1) Which of the following is not absolute measure of dispersion?

   a) Range b) Quartile deviation

   c) Stander deviation d) Coefficient of variation

2) For the data 8, 1, 4, 5, 9, 3, 2, 7 the range is ____

   a) 6 b) 8 c) 4 d) 7

3) If upper quartile and lower quartile are 90 and 45 respectively than coefficient of quartile deviation is

   a) 0.5 b) 0.67 c) 0.4 d) 0.33

4) If S.D=24 & Mean=120 then C.V is _____

   a) 12% b) 15% c) 18% d) 20%

5) _____ is a measure of dispersion.

   a) Mean b) Median c) Mode d) Range

6) If Q_1 =15, Q_3=40 then quartile deviation is :-

a) 15 b) 12.5 c) 11.6 d) 14

7) The mean and coefficient of variation are 10 and 5 respectively. The standard deviation is :-

a) 2.5 b) 6.5 c) 3.5 d) None of the above

8) Two samples A and B have the same standard deviation but the mean of A is greater than that of B the coefficient of variation of A is

a) Greater than that of B b) Less than that of B

c) Equal to that of B d) None of these

9) Measure of dispersion which is affected most by extreme observation is :-

a) Range b) Q.D c) M.D d) S.D

10) Algebraic sum of deviation from mean is

a) Positive b) Negative c) Zero d) Difference for each case

6.12 LIST OF REFERENCES

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

7

CORRELATION

Unit Structure

7.0 OBJECTIVES:

After going through this chapter you will able to know:

- Meaning of correlation and its types.
- Properties of correlation.
- Method of calculation of coefficient of correlation.

7.1 INTRODUCTION:

In the previous two chapters we have seen statistical measures used for the analysis of a univariate data, *i.e.* a data in one variable only. For example, weight, height, marks, wages, income, price *etc*. But there are sets of data which can be related to each other. Let us consider the weights and heights of people. Medically too it is said that are very closely related. In problems related to science, business and economics also it is important to know whether two variables are related to each other or not. For example, the relation between density of pollutants in air and number of vehicles, investments in advertisement and sales of a product, use of a vaccine and

number of patients, ranks or marks given by different judges in reality show, marks of same students in two consecutive exams *etc*. The question immediate is, if they have some relation then what is the magnitude and direction of that relation?

This relation between a bivariate data is said to be correlation and the magnitude of correlation is called as the correlation coefficient. The study of the magnitude and nature of correlation between two variables is called as correlation analysis.

7.2 IMPORTANCE OF CORRELATION:

(1) Correlation gives an answer to the basic question of measuring the extent of correlation between two variables.

(2) Correlation helps to understand the behavioral pattern of different variables in business and economics.

(3) The most important aspect of any analysis is decision making. Correlation finds the magnitude and direction of association between two variables and hence facilitates in the decision making process.

(4) Correlation provides a comfortable platform for estimation or forecasting, which is again an important tool in statistical analysis.

7.3 PROPERTIES OF CORRELATION COEFFICIENT:

(1) The measure of correlation is called as the coefficient of correlation and is denoted by r.

(2) It is independent of the units of measurement of the variables.

(3) The **value of *r* ranges between –1 and 1** and depends on the slope of the line passing through the values of the variables. The sign of the value of r indicates the type of correlation between the two variables. This is explained in the section **7.5 (*i*)**.

(4) The values $r = -1$ and $r = 1$, are the extreme values of correlation indicating a perfect correlation in either direction.

(5) For $0 < r < 1$, we say that there is an imperfect positive correlation. This again is classified into strong and weak positive correlation (or high and low degree positive correlation).

(6) For $-1 < r < 0$, we say that there is an imperfect negative correlation. This again is classified into strong and weak negative correlation (or high and low degree negative correlation).

(7) If $r = 0$, we say that there is no correlation or zero correlation between the two variables.

(8) It is independent of the change of scale and change of origin. If a constant is added to (or subtracted from) all the values of both the variables then also the value of r remains unchanged, this is called as change of origin. The value of r remains unchanged even if all the values of the variables are divided (or multiplied) by a constant, this is called as change of scale.

7.4 CORRELATION AND CAUSATION:

Though correlation gives the magnitude of interdependency of two variables, it does not say about the cause and effect relationship. Even if there is a high magnitude of correlation between two variables it does not necessarily mean that they are having a close relationship. For example, the increase in sales of TV sets and increase in the sales of umbrellas in a city may quantitatively show strong correlation. But the sales of TV sets may be due a pay hike or say IPL cricket matches and the sales of umbrellas due to rainy season! Such type correlation is called as nonsense correlation. The following are the causes of correlation between two variables:

(1) Influence of multiple factors: The correlation between two variables under consideration may be due the influence of multiple factors. In practice, there are third party factors which may affect the variations in the variables at the same time. For example, demand and price of a certain product may be affected by the inflation, natural calamities, economical policy, *etc*. The increase in the sales of different luxury items does not means that are actually correlated as the increase can be due to a third factor common to all, like increase in income.

(2) Mutual Influence: Both the variable may be affecting each other. In economics as we can observe that the increase in price of a commodity leads to decrease in its demand. But the relationship between price and demand is dual. So, an increase in demand may lead to increase in price also. Thus, a correlation between two variables may be due to mutual influences.

(3) Coincidence: It may happen that a pair of variables shows some correlation only by a chance and that may not be universal. Such coincidences are seen in a small sample. For example, if a small sample is taken from a village for the sales of a certain product and compared with the advertising expenditure, it may show a strong correlation. But this may be due to monopoly of that product in the area from where the sample is taken and may not be true for a wider area. Thus, a conclusion that increase in advertising expenditure has lead to increase in sales, on the basis of the strong correlation may be far from true.

7.5 TYPES OF CORRELATION:

Correlation can be classified into three types as follows:

(*i*) Positive or negative Correlation:

Positive (or negative) correlation is related to the direction of the Correlation. If the increase (or decrease) in one variable results in increase (or decrease) in the other variable then we say that the correlation is *positive*. The value of r in such type of correlation is between 0 and 1. If the increase (or decrease) in one variable results in decrease (or increase) in the other variable then we say that the

correlation is *negative.* The value of r in such type of correlation is between –1 and 0.

(*ii*) Simple, Partial or multiple Correlation:

If the correlation is only between two variables then we say it as a *simple correlation*. If the other influencing factors affecting the two variables are assumed to be constant then it is said to be *partial correlation.* If more than two variables are studied for their inter dependencies then it is said to be *multiple correlation.*

(*iii*) Linear and Non-Linear Correlation:

If the relation between the two variables is linear we say it is a *linear correlation*. The meaning of being linear is that a mathematical relation of the type $y = ax + b$, which is an equation of straight line, can be established between the two variables. In other words, the values of the variables must be in constant ratio.

For example, consider the following data of two variables price and demand:

Price (x)	:	2	4	6	8	10	12
Demand (y)	:	8	12	16	20	24	28

The relation between price and demand can be written as $y = 2x + 4$.

Practically we do not get a linear relationship always. Such type of Correlation is said *non-linear correlation*. The graph of the values of the variables is not a straight line, but a curve.

There are different methods to measure the magnitude or to understand the extent of correlation between two variables. In this chapter we are going to study three such methods:

(1) Scatter Diagram, (2) Correlation Table, (3) Correlation Graph and (4) Correlation Coefficient. As per our scope of syllabus, we shall be studying in detail two types of Correlation Coefficients namely, Karl Pearson's Coefficient of Correlation and Spearman's rank Correlation Coefficient.

7.6 SCATTER DIAGRAM:

This is the simplest method to study the correlation between two variables. *Scatter Diagram* is a graphical method to study the extent of correlation between variables. A graph of the two variables X and Y is drawn by taking their values on the corresponding axis. Points are plotted on the graph and the conclusion is made on the density of the points on the graph. The following figures and the explanations would make it clearer.

(*i*) Perfect Positive Correlation:

If the graph of the values of the variables is a straight line with positive slope as shown in Figure **7.1**, we say there is a *perfect positive correlation* between *X* and *Y*. Here $r = 1$.

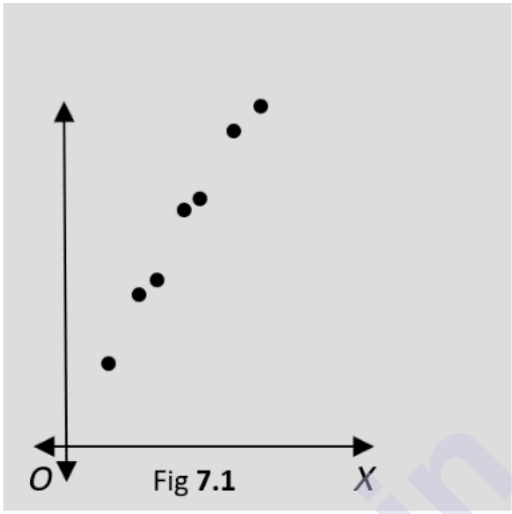


Fig **7.1**

(*ii*) Imperfect Positive Correlation:

If the graph of the values of *X* and *Y* show a band of points from lower left corner to upper right corner as shown in Figure **7.2**, we say that there is an *imperfect positive correlation*. Here $0 < r < 1$.

(*iii*) Perfect Negative Correlation:

If the graph of the values of the variables is a straight line with negative slope as shown inFigure **7.3**, we say there is a *perfect negative correlation* between *X* and *Y*. Here $r = -1$.

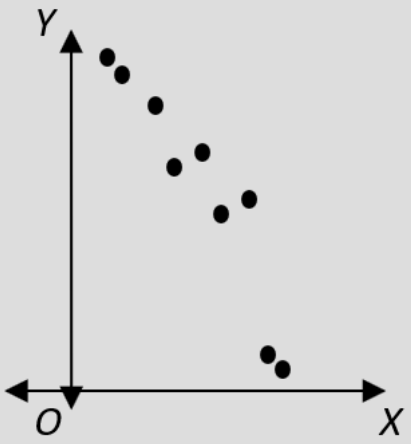


Fig. **7.3**

(*iv*) Imperfect Negative Correlation:

If the graph of the values of *X* and *Y* show a band of points from upper left corner to the lower right corner as shown in Figure **7.4**, then we say that there is an *imperfect negative correlation*. Here $-1 < r < 0$

(*v*) Zero Correlation:

If the graph of the values of *X* and *Y* do not show any of the above trend then we say that there is a zero correlation between *X* and *Y*. The graph of such type can be a straight line perpendicular to the axis, as shown in Figure **7.5** and **7.6,** or may be completely scattered as shown in Figure **7.7**. Here $r = 0$.

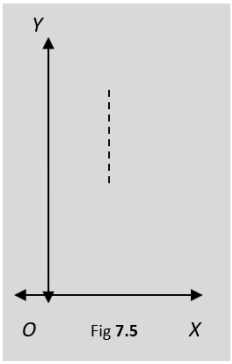


Fig 7.5

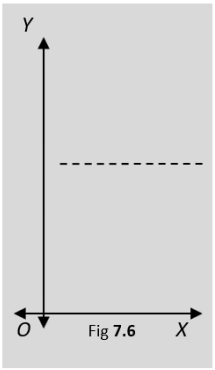


Fig 7.6

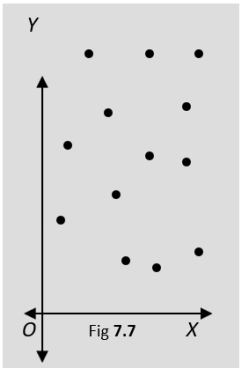


Fig 7.7

The Figure **7.5** show that the increase in the values of *Y* has no effect on the value of *X*, it remains the same, hence zero correlation. The Figure **7.6** show that the increase in the values of *X* has no effect on the value of *Y*, it remains the same, hence zero correlation. The Figure **7.7** show that the points are completely scattered on the graph and show no particular trend, hence there is no correlation or zero correlation between *X* and *Y*.

7.6.1 Merits and demerits of scatter diagram:

Merits

(1) It is very easy to understand and interpret the degree of correlation.

(2) It is a simple method and involves no mathematical calculations.

(3) It is not affected by extreme values.

Demerits

(1) It fails to give the exact magnitude of correlation.

(2) It cannot be used for any further analysis.

7.7 KARL PEARSON'S COEFFICIENT OF CORRELATION:

The Karl Pearson's Correlation Coefficient is also known as the *product moment correlation coefficient.* The coefficient of correlation as developed by Karl Pearson is defined as the ratio of the Covariance between *x* and *y* to the product of respective standard deviations. Thus,

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y} \qquad \ldots (1)$$

where *Cov*(*x*, *y*) is the Covariance between *x* and *y* and σ_x, σ_y are the standard deviations of *x* and *y* respectively.

Covariance between *x* and *y*, which is the average of sum of the product of deviations of the values from their respective averages, is given by the

formula: $Cov(x,y)=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$ and we know that: $\sigma_x=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$, $\sigma_y=\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$

Substituting in (1), we have $r=\frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}}=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$

.... (2)

The simplified form of the above formula is as follows:

$$r=\frac{\Sigma xy-\frac{\Sigma x\Sigma y}{n}}{\sqrt{\Sigma x^2-\frac{(\Sigma x)^2}{n}}\sqrt{\Sigma y^2-\frac{(\Sigma y)^2}{n}}} \quad \ldots (3)$$

or $$r=\frac{n\Sigma xy-\Sigma x\Sigma y}{\sqrt{n\Sigma x^2-(\Sigma x)^2}\sqrt{n\Sigma y^2-(\Sigma y)^2}} \quad \ldots (4)$$

The formula (3) or (4) is the most commonly used formula for computing the correlation coefficient from a raw data. If the sum of product of deviations and the standard deviations are known directly then formula (1) or (2) can be used.

Steps to find Karl Pearson's Correlation Coefficient '*r*' from a raw data:

(1) The columns of xy, x^2 and y^2 are introduced.

(2) The respective column totals: $\Sigma x, \Sigma y, \Sigma xy, \Sigma x^2$ and Σy^2 are calculated.

(3) Using the above formula no. (3) or (4), r is calculated.

Note: The value of r is between – 1 and 1, if we get a value which is not in this range it means our calculations are not correct!

Example 1: Calculate the Karl Pearson's coefficient of correlation for the following data and comment:

x	10	8	11	7	9	12
y	8	5	10	6	7	11

Solution: Introducing the columns as mentioned above, the table of computation is:

x	y	xy	x^2	y^2
10	8	80	100	64
8	5	40	64	25

11	10	110	121	100
7	6	42	49	36
9	7	63	81	49
12	11	131	144	121
$\Sigma x = 57$	$\Sigma y = 47$	$\Sigma xy = 466$	$\Sigma x^2 = 559$	$\Sigma y^2 = 395$

Here $n = 6$. Using the calculated totals from the table and the formula no. (4), we have

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{6 \text{ x } 466 - 57 \text{ x } 47}{\sqrt{6 \text{ x } 559 - (57)^2}\sqrt{6 \text{ x } 395 - (47)^2}}$$

(Students should be careful in substituting and calculating the values properly in the above formula. Especially in the case of Σx^2 and $(\Sigma x)^2$. Neither Σx^2 **is not the square of** Σx as can be seen from the table nor $(\Sigma x)^2$ is same as Σx^2 !!)

$$\therefore r = \frac{2796 - 2679}{\sqrt{3594 - 3249}\sqrt{2370 - 2209}} = \frac{117}{\sqrt{345}\sqrt{161}} = \frac{117}{18.57 \text{ x } 12.68}$$

$$\therefore r = 0.49$$

Since $r = 0.49 > 0$, there is an imperfect positive correlation between x and y. (to be more precise *weak imperfect positive correlation*)

Example 2: Calculate the coefficient of correlation from the following given information and comment: $n = 10$, $\Sigma x = 608$, $\Sigma y = 640$, $\Sigma xy = 39965$, $\Sigma x^2 = 39054$ and $\Sigma y^2 = 42096$

Solution: All the required totals are provided, hence using the formula no. (4), we have

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{10 \text{ x } 39965 - 608 \text{ x } 640}{\sqrt{10 \text{ x } 39054 - (608)^2}\sqrt{10 \text{ x } 42096 - (640)^2}}$$

$$\therefore r = \frac{399650 - 389120}{\sqrt{390540 - 369664}\sqrt{420960 - 409600}} = \frac{10530}{\sqrt{20876}\sqrt{11360}} = \frac{10530}{144.48 \text{ x } 106.58}$$

$$\therefore r = 0.68$$

Thus, there is an imperfect positive correlation between x and y.

Example 3: Calculate the coefficient of correlation from the following given information and comment: $n = 6$, $\Sigma x = 105$, $\Sigma y = 305$, $\Sigma xy = 5110$, $\Sigma x^2 = 1855$ and $\Sigma y^2 = 18525$

Solution: All the required totals are provided, hence using the formula no. (4), we have

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{6 \text{ x } 5110 - 105 \text{ x } 305}{\sqrt{6 \text{ x } 1855 - (105)^2}\sqrt{6 \text{ x } 18525 - (305)^2}}$$

$$\therefore r = \frac{30660 - 32025}{\sqrt{11130 - 11025}\sqrt{111150 - 93025}} = \frac{-1365}{\sqrt{105}\sqrt{18125}} = \frac{-1365}{10.25 \text{ x } 134.63}$$

$$\therefore r = -0.98$$

Since, r = -0.98 < 0, there is a strong imperfect negative correlation.

Example 4: Calculate the coefficient of correlation from the following given information and comment: $\Sigma(x-\bar{x})(y-\bar{y}) = 1240$, $\Sigma(x-\bar{x})^2 = 1650$ and $\Sigma(y-\bar{y})^2 = 2430$.

Solution: From the formula no. (2), we know that: $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$

$$\therefore r = \frac{1240}{\sqrt{1650}\sqrt{2430}} = \frac{1240}{40.62 \text{ x } 49.29} = 0.62$$

Thus, there is imperfect positive correlation.

Example 5: Calculate the product moment correlation coefficient and comment:

x	75	60	55	50	48	45
y	70	80	82	85	90	94

Solution: In problems where the values of the variables are large, change of origin can be done to simplify the problem. We know from property (8) of r that it is independent of change of origin and scale.

Let us assume $u = x - 55$ and $v = y - 80$. We know by above mentioned property that, $r_{uv} = r_{xy}$. Now introducing columns of the type uv, u^2 and v^2, we prepare the table of calculation as follows:

x	y	u	v	Uv	u^2	v^2
75	70	20	-10	-200	400	100
60	80	5	0	0	25	0
55	82	0	2	0	0	4
50	85	-5	5	-25	25	25
48	90	-7	10	-70	49	100

45	94	-10	14	-140	100	196
Total		$\Sigma u =$ 3	$\Sigma v =$ 21	$\Sigma uv =$ -435	$\Sigma u^2 =$ 599	$\Sigma v^2 =$ 425

Here, $n = 6$.The change of origin formula is given as:

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

$$\therefore r_{uv} = \frac{6 \text{ x } (-435) - 3 \text{ x } 21}{\sqrt{6 \text{ x } 599 - (3)^2}\sqrt{6 \text{ x } 425 - (21)^2}}$$

$$\therefore r_{uv} = \frac{-2610 - 63}{\sqrt{3594 - 9}\sqrt{2550 - 441}} = \frac{-2673}{\sqrt{3585}\sqrt{2109}} = \frac{-2673}{59.87 \text{ x } 45.92}$$

$$\therefore r_{uv} = -0.97 = r_{xy}$$

Thus, there is a strong negative correlation between x and y.

Example 6: Calculate the Karl Pearson's correlation coefficient and comment:

x	100	150	200	300	400	550
y	20	30	40	50	60	70

Solution: If the values of the variables are multiples of a common number then the problem can be simplified by using change of scale, as we know that the correlation coefficient is independent of change of scale.

Let $u = \frac{x - 300}{50}$ and $v = \frac{y - 40}{10}$. Now introducing columns of the type uv, u^2 and v^2, we prepare the table of calculation as follows:

x	y	u	v	Uv	u^2	v^2
100	20	-4	-2	8	16	4
150	30	-3	-1	3	9	1
200	40	-2	0	0	4	0
300	50	0	1	0	0	1
400	60	2	2	4	4	4
550	70	5	3	15	25	9
Total		$\Sigma u =$ - 2	$\Sigma v =$ 3	$\Sigma uv =$ 30	$\Sigma u^2 =$ 58	$\Sigma v^2 =$ 19

Here $n = 6$. Since $r_{uv} = r_{xy}$, the formula and the calculations are as follows:

$$r_{uv} = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}} = \frac{6 \text{ x } (30) - (-2) \text{ x } 3}{\sqrt{6 \text{ x } 54 - (-2)^2}\sqrt{6 \text{ x } 19 - (3)^2}}$$

$$\therefore r_{uv} = \frac{180 + 6}{\sqrt{348 - 4}\sqrt{114 - 9}} = \frac{186}{\sqrt{344}\sqrt{105}} = \frac{186}{18.55 \text{ x } 10.25}$$

$$\therefore r_{xy} = r_{uv} = 0.978$$

Thus, there is a strong positive correlation between x and y.

7.7.1 Coefficient of correlation for a grouped data:

If the values of the variables are given as a grouped data with their corresponding frequencies, a bivariate table is needed to be calculated. The formula for calculation is as follows:

$$r = \frac{N\Sigma fxy - \Sigma fx \Sigma fy}{\sqrt{N\Sigma fx^2 - (\Sigma fx)^2}\sqrt{N\Sigma fy^2 - (\Sigma fy)^2}} \quad \ldots$$

(5)

This formula is similar to what we have been using till now. The difference in notations is that now the corresponding frequencies are getting multiplied to each term in the formula.

But the steps to calculate the correlation coefficient have to be very carefully understood as the bivariate table has class intervals and frequencies of both the variables.

Note: While practicing the solved problem, students should refer to the steps given below, one by one and then compare it with the calculations done in the table.

Steps to find correlation coefficient for a Grouped Bivariate data

(1) Given a bivariate grouped data, we introduce the column and row of their mid points (say x and y). If the data is discrete (*i.e.* without class intervals) this step is to be skipped as the values of x and y are already given. If change of scale or origin is needed, it should be done in this step.

(2) Total the frequencies in each column and in each row, write them in a new column and row named as f. The horizontal and vertical total of these column values and row values is same and is denoted as N.

(3) Multiply the frequency in each cell with its corresponding values of x and y and write it in parenthesis inside the same cell. These values can also be put in a box or circled to differentiate them from the frequency in that cell.

(4) Now introduce the columns and rows of fx, fx^2, fxy and fy, fy^2, fxy (in case of change of origin or scale these columns and rows are fu, fu^2, fuv and fv, fv^2, fuv)

(5) The values in the cells under fx are found by multiplying the total frequency in that column (or row) with the corresponding mid points. The values of fx^2 are found by multiplying the values of fx with x. This step has to be repeated for finding values for fy and fy^2. The totals of all columns and rows of fx, fx^2, fy, fy^2 are to be calculated.

(6) Now, the values of entries in the column (and row) of fxy are the respective totals of the values written in the upper right corner of every cell. The horizontal and vertical total of all fxy's should be same.

(7) The totals thus obtained are substituted in the formula no. (5) and the correlation coefficient is finally computed!

Solved below are three different examples covering different cases of bivariate grouped data.

Example 7: Find the correlation coefficient for the following bivariate grouped data:

	y			
x	2	4	6	8
1	-	3	5	1
3	1	-	-	4
5	4	-	2	-
7	-	1	-	-

Solution: The given bivariate data is a discrete data. No entry in any cell means the frequency for that cell is 0. Skipping the first step, we calculate the row and column totals of the corresponding frequencies as follows:

	y				
x	2	4	6	8	f
1	0	3	5	1	0 +3+5+1 = 9
3	1	0	0	4	1+0+0+4 = 5
5	4	0	2	0	4+0+2+0 = 6
7	0	1	0	0	0+1+0+0 = 1
f	5	4	7	5	$N = 21$

These are the column totals of the

These are the row totals of the frequencies from each cell.

Observe that the total frequency N is same horizontally and vertically.

Now, using step (3), the frequencies in each cell are multiplied with orresponding values of x and y and the product is written in the parenthesis inside the cell as shown below.

The value in the shaded cell is calculated by multiplying 4 with 5 and 2,*i.e.* 4 x 5 x 2 = 40

		Y			
x	2	4	6	8	*F*
1	0 (0)	3 (12)	5 (30)	1 (8)	9
3	1 (6)	0 (0)	0 (0)	**4 (96)**	5
5	**4 (40)**	0 (0)	2 (60)	0 (0)	6
7	0 (0)	1 (28)	0 (0)	0 (0)	1
f	5	4	7	5	*N*=21

The value in the shaded cell is calculated by multiplying 4 with 3 and

Now, the final table is completed by introducing the columns of fx, fx^2 and fxy and the rows of fy, fy^2 and fxy as shown below.

	Y							
x	2	4	6	8	*F*	*fx*	fx^2	*fxy*
1	0 (0)	3 (12)	5 (30)	1 (8)	9	9	9	50
3	1 (6)	0 (0)	0 (0)	4 (96)	5	15	45	102
5	4 (40)	0 (0)	2 (60)	0 (0)	6	30	150	100
7	0 (0)	1 (28)	0 (0)	0 (0)	1	7	49	28
f	5	4	7	5	$N = 21$	Σfx =61	Σfx^2 =253	Σfxy =**280**
fy	10	16	42	40	Σfy = 108			
fy^2	20	64	252	320	Σfy^2 = 656			
fxy	46	40	90	104	Σfxy = **280**			

From the table the values of the required terms are:

$N = 21$, $\Sigma fx = 61$, $\Sigma fx^2 = 253$, $\Sigma fy = 108$, $\Sigma fy^2 = 656$ and $\Sigma fxy = 280$.

$$r = \frac{N\Sigma fxy - \Sigma fx\Sigma fy}{\sqrt{N\Sigma fx^2 - (\Sigma fx)^2}\sqrt{N\Sigma fy^2 - (\Sigma fy)^2}} = \frac{21 \text{ x } 280 - 61 \text{ x } 108}{\sqrt{21 \text{ x } 253 - (61)^2}\sqrt{21 \text{ x } 656 - (108)^2}}$$

$$\therefore r = \frac{-708}{\sqrt{1592}\sqrt{2112}} = \frac{-708}{39.89 \text{ x } 45.96}$$

$$\therefore r = -0.386$$

Example 8: Calculate the Karl Pearson's correlation coefficient for the following data:

Weight in kg / Height in cm	5 – 15	15 – 25	25 – 30
40 – 60	10	5	-
60 – 80	5	8	2
80 – 100	-	10	12

Solution: The mid points of class intervals are found in the first step and the row and column totals of the frequencies are completed as follows:

y / x	10	20	30	F
50	10	5	0	15
70	5	8	2	15
90	0	10	12	22
f	15	23	14	$N = 52$

Now the procedure as done in above example is followed and the table is completed as shown below:

x \ y	10	20	30	F	fx	fx^2	fxy
50	10 (5000)	5 (5000)	0 (0)	15	750	37500	10000
70	5 (3500)	8 (11200)	2 (4200)	15	1050	73500	18900
90	0 (0)	10 (18000)	12 (32400)	22	1980	178200	50400
F	15	23	14	$N=52$	3780	289200	**79300**
Fy	150	460	420	1030			
fy^2	1500	9200	12600	23300			
fxy	8500	34200	36600	**79300**			

From the table the values of the required terms are:

$N = 52$, $\Sigma fx = 3780$, $\Sigma fx^2 = 289200$, $\Sigma fy = 1030$, $\Sigma fy^2 = 23300$ and $\Sigma fxy = 79300$.

$$r = \frac{N\Sigma fxy - \Sigma fx \Sigma fy}{\sqrt{N\Sigma fx^2 - (\Sigma fx)^2}\sqrt{N\Sigma fy^2 - (\Sigma fy)^2}}$$

$$= \frac{52 \text{ x } 79300 - 3780 \text{ x } 1030}{\sqrt{52 \text{ x } 289200 - (3780)^2}\sqrt{52 \text{ x } 23300 - (1030)^2}}$$

$$\therefore r = \frac{230200}{\sqrt{750000}\sqrt{150700}} = \frac{230200}{388.2 \text{ x } 866.03} = 0.6847$$

Example 9: Compute the coefficient of correlation, for the following data:

Group I \ Group II	20 – 40	40 – 60	60 – 80
0 – 10	5	7	4
10 – 20	2	6	3
20 – 30	4	2	7

Solution: The previous problem was solved by direct method. But this problem we shall solve using the change of scale method. The mid points of the Group I are 5, 15 and 25, so we assume u as

$u = \frac{x-15}{10}$. The mid points of Group II are 30, 50 and 70, so let us assume $v = \frac{y-50}{20}$.

The initial table can be completed as follows:

Group II			20 – 40	40 – 60	60 – 80	
Mid points y			30	50	70	
Group I	Mid points x	$u = \frac{x-15}{10}$ \ $v = \frac{y-50}{20}$	– 1	0	1	f
0 – 10	5	– 1	5	7	4	16
10 – 20	15	0	2	6	3	11
20 – 30	25	1	4	2	7	13
		f	11	15	14	$N = 40$

Now we the table is completed by following the remaining steps as shown below. The initial two rows and columns are not shown below as we have changed the scale.

v \ u	-1	0	1	F	fu	fu^2	fuv
-1	5 (5)	7 (0)	4 (-4)	16	-16	16	1
0	2 (0)	6 (0)	3 (0)	11	0	0	0
1	4 (-4)	2 (0)	7 (7)	13	13	13	3
F	11	15	14	$N = 40$	$\Sigma fu = -3$	$\Sigma fu^2 = 29$	$\Sigma fuv = \mathbf{4}$
Fv	-11	0	14	$\Sigma fv = 3$			
fv^2	11	0	14	$\Sigma fv^2 = 25$			
fuv	1	0	3	$\Sigma fuv = \mathbf{4}$			

From the table the values of the required terms are substituted in the formula as shown below:

$$r_{uv} = \frac{N\Sigma fuv - \Sigma fu\Sigma fv}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\sqrt{N\Sigma fv^2 - (\Sigma fv)^2}} = \frac{40 \text{ x } 4 - (-3) \text{ x } 3}{\sqrt{40 \text{ x } 29 - (-3)^2}\sqrt{40 \text{ x } 25 - (3)^2}}$$

$$\therefore r_{uv} = \frac{160 + 9}{\sqrt{1160 - 9}\sqrt{1000 - 9}} = \frac{169}{\sqrt{1151}\sqrt{991}} = \frac{169}{33.93 \text{ x } 31.48}$$

$$\therefore r_{xy} = r_{uv} = 0.158$$

7.7.2 Merits and demerits of karl pearson's coefficient of correlation

Merits

(1) It gives the magnitude and direction of the correlation between two variables.

(2) It is the most commonly used and popular measure of finding correlation coefficient.

Demerits

(1) The formula and method is not very easy to remember and understand quickly.

(2) It does not ascertain the existence of correlation between two variables. In other words, there may not be any actual relationship between variables but the value of *r* may not say so. The correlation coefficient has to does not give the cause and effect relationship between variables. There is thus, a high chance of misinterpretation.

(3) It is affected by extreme values of the variables taken into consideration.

(4) It cannot be used for a non linear relationship. The formula assumes that there is always a linear relation between the variables, whereas actually it may not be so.

7.8 SPEARMAN'S RANK CORRELATION COEFFICIENT

This formula developed by Charles Spearman is useful in measuring the correlation between two variables when the data is given in a certain order. This order is generally ranks given to the variables based on some qualitative information. Fir example, ranks to students based on their performance, ranks to the contestants of some competition, ranks to TV serials based on their TRP's *etc*.

The Spearman's Rank Correlation Coefficient is denoted by *R*. In general the formula is as given below:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad \ldots (6)$$

Where n: number of observations

d^2: The square of differences between the ranks

If R_1 denotes ranks given to the first variable, R_2 denotes the ranks given to the second variable, then $d = R_1 - R_2$. Every such difference is squared and finally totaled to get Σd^2.

The value of R, the Spearman's rank correlation coefficient, like that of the Karl Pearson's correlation coefficient also ranges between – 1 and 1.

There are three different types of problems related to the formula. We shall discuss each separately and understand it with suitable examples.

If Ranks are given directly: If the ranks to the variables are already given then it is very simple to compute the value of R.

Step IFor every observation the difference between the ranks, *i.e.* $d = R_1 - R_2$ is calculated.

Step II The column of the squares of these differences is introduced and its sum Σd^2 is calculated.

Step III The formula no. (6) mentioned above is used to compute the rank correlation coefficient.

Example 10: The following data gives the ranks of 10 students in two consecutive years 1990 and 1991

1990	2	4	1	7	3	9	6	10	8	5
1991	3	2	1	5	6	7	8	9	10	4

Find the rank correlation coefficient.

Solution: Let the ranks in the year 1990 be denoted by R_1 and those for the year 1991 be denoted by R_2. The above mentioned steps are followed and the table of calculations is completed as follows:

R_1	R_2	$d = R_1 - R_2$	d^2
2	3	-1	1
4	2	2	4
1	1	0	0
7	5	2	4
3	6	-3	9
9	7	2	4
6	8	-2	4
10	9	1	1
8	10	-2	4
5	4	1	1
	$n = 10$		$\Sigma d^2 = 32$

The Spearman's rank correlation coefficient is calculated as follows:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 32}{10(10^2-1)}$$ (*students should not square* 32 *here*!!)

$$\therefore R = 1 - \frac{192}{990} = 1 - 0.1939 = 0.808$$

Thus, there is strong positive correlation between the two performances in the two years.

Example 11: In a singing contest twelve participants were judged by three judges. The ranks given to the participants by the judges are given below. Find which pair of judges have a most common approach in judgment.

Judge A	2	7	6	3	1	9	11	8	12	4	10	5
Judge B	5	4	7	1	2	6	8	12	11	3	9	10
Judge C	1	6	7	3	2	8	12	9	11	5	10	4

Solution: In such kind of problems we consider the data of the variables in pairs. Here we will be finding the rank correlation between Judge A and B, Judge A and C and Judge A and C.

For Judge A and B				For Judge A and C				For Judge B and C			
R_A	R_B	$d_{AB} = R_A - R_B$	d_{AB}^2	R_A	R_C	$d_{AC} = R_A - R_C$	d_{AC}^2	R_B	R_C	$d_{BC} = R_B - R_C$	d_{BC}^2
2	5	-3	9	2	1	1	1	5	1	4	16
7	4	3	9	7	6	1	1	4	6	-2	4
6	7	-1	1	6	7	-1	1	7	7	0	0
3	1	2	4	3	3	0	0	1	3	-2	4
1	2	-1	1	1	2	-1	1	2	2	0	0
9	6	3	9	9	8	1	1	6	8	-2	4
11	8	3	9	11	12	-1	1	8	12	-4	16
8	12	-4	16	8	9	-1	1	12	9	3	9
12	11	1	1	12	11	1	1	11	11	0	0
4	3	1	1	4	5	-1	1	3	5	-2	4
10	9	1	1	10	10	0	0	9	10	-1	1
5	10	-5	25	5	4	1	1	10	4	6	36
		$\Sigma d_{AB}^2 = 86$				$\Sigma d_{AC}^2 = 10$				$\Sigma d_{BC}^2 = 94$	

Here $n = 12$.

Here n = 12.

For Judge A and B: $\Sigma d_{AB}^2 = 86$

$$R_{AB} = 1 - \frac{6\Sigma d_{AB}^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 86}{12(12^2-1)} = 1 - \frac{516}{1716} = 1 - 0.3$$

$\therefore R_{AB} = 0.7$ … (i)

For Judge A and C: $\Sigma d_{AC}{}^2 = 10$

$$R_{AC} = 1 - \frac{6\Sigma d_{AC}{}^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 10}{12(12^2-1)} = 1 - \frac{60}{1716} = 1 - 0.035$$

$\therefore R_{AC} = 0.965$ … (ii)

For Judge B and C: $\Sigma d_{BC}{}^2 = 94$

$$R_{BC} = 1 - \frac{6\Sigma d_{BC}{}^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 94}{12(12^2-1)} = 1 - \frac{564}{1716} = 1 - 0.3286$$

$\therefore R_{BC} = 0.6714$ … (iii)

Comparing the values of the rank coefficients (i), (ii) and (iii), we observe that the value of rank

correlation between Judge A and C is the highest, among the three pairs.

Thus, Judge A and Judge C have the most common approach in judgment.

7.8.1 Ranks are not given and are non – repeated ranks

If the instead of ranks the actual data is given related to the variables we rank the data. The ranks can be given in any order either ascending or descending. The highest value gets the rank 1, and so on till the smallest value getting the n^{th} rank. Once the ranks are given, the remaining steps are similar to the previous type. In this sub section we shall see example where the values are not repeated.

Example 12: Find the rank correlation coefficient for the following data giving marks of FYBMS students in the subjects of Mathematics and Statistics:

Marks in Maths	65	45	78	35	52	73	67	49	40
Marks in Stats	60	40	70	28	72	59	69	56	55

Solution: The marks in the subjects are ranked with the highest getting 1st rank and the least one getting 9th rank. The table of computation is further completed as shown below:

Marks in Maths	R_1	Marks in Stats	R_2	d	d^2
65	4	60	4	0	0
45	7	40	8	-1	1
78	1	70	2	-1	1

	35	9	28	9	0	0
	52	5	72	1	4	16
	73	2	59	5	-3	9
	67	3	69	3	0	0
	49	6	56	6	0	0
From the table: n = 9 and $\Sigma d^2 = 28$	40	8	55	7	1	1
			$n = 9$			$\Sigma d^2 = 28$

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 28}{9(9^2-1)} = 1 - \frac{168}{720} = 1 - 0.23$$

$$\therefore R = 0.77$$

Thus, there is a imperfect positive correlation between the marks in both the subjects.

7.8.2 Repeated ranks

If two or more observations in a data set have same values then the ranks cannot be discrete. To avoid this, such observations are given rank equal to the arithmetic mean of the ranks that would have been given if they were different and in order.

Let us consider an example of a data related to marks of students. After giving first two ranks if there are three students with same marks, then each is given rank equal to the average of 3, 4 and 5. If they would had been distinct and in order, then each would had got one of these ranks. Thus, each observation here gets the rank 4 which is the average of 3.

If after giving first five ranks there are two students with same marks, then again both of them are given rank equal to the mean of the next two ranks *i.e.* mean of 6 and 7, which is 6.5. Thus, each observation in this case gets rank 6.5

In formulating the formula for rank correlation it is assumed that each observation is ranked distinctly. When repeated ranks are assigned a correction factor (C.F.) is to be calculated for every repeated rank and their total called as T.C.F., total correction factor is added to the Σd^2 value in the actual formula. The correction factor for a particular rank is calculated by the formula: $\text{C.F.} = \frac{m(m^2-1)}{12}$, where m is the number of times that rank is repeated. In above example, in the first case m = 3. $\therefore \text{C.F.} = \frac{3(3^2-1)}{12} = \frac{24}{12} = 2$. In the second case m = 2. $\therefore \text{C.F.} = \frac{2(2^2-1)}{12} = \frac{6}{12} = 0.5$. Hence T.C.F. = 2 + 0.5 = 2.5

The new formula of rank correlation for repeated ranks is: $R = 1 - \frac{6(\Sigma d^2 + \text{TCF})}{n(n^2 - 1)}$

Note: In the numerator of the formula, the TCF is added to Σd^2 and their total is multiplied with 6.

Example 13: Calculate the rank correlation coefficient for the following data:

Data I	10	7	12	10	16	14	5	11	18	7
Data II	6	11	5	9	6	10	15	18	6	8

Solution: Let us observe that in the first data, 10 is repeated two time and 7 is also repeated two times. In the second data, 6 is repeated three times.

Following the procedure explained in the above section, the table of calculations is completed as shown below:

Data I	R_1	Data II	R_2	$d = R_1 - R_2$	d^2
10	6.5	6	8	-1.5	2.25
7	8.5	11	3	5.5	30.25
12	4	5	10	-6	36
10	6.5	9	5	1.5	2.25
16	2	6	8	-6	36
14	3	10	4	-1	1
5	10	15	2	8	64
11	5	18	1	4	16
18	1	6	8	-7	49
7	8.5	8	6	2.5	6.25
		n = 10			Σd^2 = 243

Before proceeding with the further calculations let us understand how the ranks are given to the values in Data I and Data II.

In Data I, the first five ranks are in order. The next value 10 is repeated twice and is given rank equal to the mean of the next two ranks *i.e.* of 6 and 7. The mean of 6 and 7 is 6.5, so rank 6.5 is given to both the values. The next value below 10 is 7 which is also repeated twice. The next two ranks

in the order now are 8 and 9 (as 6 and 7 have been utilized). Thus, these values get a rank equal to the mean of 8 and 9 which is 8.5.

In Data II, the first six ranks are in order. The next lower value is 6 which is repeated thrice. Hence it is given a rank equal to the mean of the next three ranks which are 7, 8 and 9. Hence, these values are ranked as 8, which is the average of 7, 8 and 9.

Calculation of Correction Factor:

For Data I: (i) 10 is repeated two times, so $m = 2$. $\therefore$ CF = 0.5

(ii) 7 is repeated two times, so $m = 2$. $\therefore$ CF = 0.5

For Data II:(i) 6 is repeated three times, so $m = 3$. $\therefore$ CF = 2

$\therefore$ TCF = 0.5 + 0.5 + 2 = 3

Now, $n = 10$, $\Sigma d^2 = 243$ and TCF = 3

$$R = 1 - \frac{6(\Sigma d^2 + \text{TCF})}{n(n^2 - 1)} = 1 - \frac{6 \text{ x } (243+3)}{10(10^2 - 1)} = 1 - \frac{1476}{990} = 1 - 01.49$$

$\therefore R = -0.49$

Thus, there is a negative correlation between the given two sets of data.

Example 14: If the sum of the squares of differences in the ranks of two variables is 82.5 and the rank correlation coefficient is 0.5, find the number of observations.

Solution: Given: $R = 0.5$ and $\Sigma d^2 = 82.5$

We know that $R = 1 - \frac{6(\Sigma d^2)}{n(n^2 - 1)}$. Substituting the values given we get,

$$0.5 = 1 - \frac{6(82.5)}{n(n^2 - 1)} = 1 - \frac{495}{n(n^2 - 1)}$$

$$\therefore \frac{495}{n(n^2 - 1)} = 1 - 0.5 = 0.5$$

$$\therefore n(n^2 - 1) = \frac{495}{0.5} = 990 = 10 \text{ x } 99 = 10\,(10^2 - 1)$$

$\therefore n = 10$.

7.9 LET US SUM UP

In this chapter we have learn:

- Definition of correlation and its types.
- Properties of correlation.
- Graphical method of correlation to explain types of correlation.

- Karl Pearson's method to calculate coefficient of correlation.
- Spearman's rank correlation.

7.10 UNIT END EXERCISES

1. Define Correlation and coefficient of correlation.
2. Explain the importance of Correlation.
3. Describe the different types of correlation with suitable diagrams and examples.
4. Interpret the values of *r*.
5. What are different types of correlation coefficients?
6. Discuss the merits and demerits of correlation coefficients.
7. Compute the coefficient of correlation for the following data:

X	7	9	8	5	6	3	4	1	2
Y	18	20	19	21	24	26	25	23	27

8. Compute the coefficient of correlation for the following data:

X	90	102	106	110	120	115	119	114	111
Y	60	70	64	67	69	73	71	68	66

9. Find the coefficient of correlation for the data given below representing the income and expenditure of families in a city. The income and expenditure values are in '000 Rs.:

Income	7	9	11	13	15	17
Expenditure	4	5	7	10	13	16

10. The following data gives the percentage of students in their SSC and HSC examination. Find the coefficient of correlation.

SSC	60	55	48	72	81	88	59	42	77
HSC	55	54	50	78	80	86	67	48	74

11. The following data gives the details of imports and exports in terms of money (in lakhs of Rs.) for a country. Find the coefficient of correlation.

Imports	22	25	21	26	29	31	28	32	33	35
Exports	36	38	40	46	42	40	44	50	52	57

12. The following data gives marks obtained by 8 students in two tests. Find the coefficient of correlation between the performances in two tests.

Test I	5	17	13	8	22	18	12	14
Test II	8	20	12	5	18	20	10	11

If the class teacher later on gives 5 marks for attendance to all the students, what will be the correlation coefficient then?

13. The following data gives the sensitive index number in BSE and NSE for 5 consecutive days. Find the correlation coefficient between them.

BSE	15540	15600	16022	14870	14900	15120
NSE	652	660	710	590	625	634

14. The following data gives the average rainfall (in cm) in an area and yield of a crop (in tons). Find the correlation coefficient between them. Are the two related to each other?

Rainfall	120	168	170	165	150	172	180	175
Yield	79	82	90	121	156	175	190	230

15. The following data gives the heights of father and their sons. Find whether there is any correlation between the two heights?

Height of father	162	156	166	172	171	178	153	160	177	180
Height of Son	156	160	155	170	165	168	160	154	180	182

16. The following data give the amount of chemical fertilizer (X) used by a farmer over a period of 10 years, the yield of the crop (Y) and the percentage of minerals (Z) in the farm. Find the correlation between X and Y, X and Z.

X	10	12	14	16	18	20	22	24	26	28
Y	88	90	96	102	110	109	108	105	106	104
Z	70	68	66	65	64	62	61	60	58	56

17. The following data gives the amount of pocket money given to college going students and their expenditure on eatables. Find the correlation coefficient.

Money	100	150	200	250	300	350
Expenditure on eatables	45	80	120	150	180	230

18. The following data gives the percentage of toppers in their SSC and HSC examination. Find the degree of correlation is there between the two results.

SSC	90	92	91	96	97	95
HSC	86	90	85	92	91	70

19. Raut Pharma ltd. wants to know about the impact of advertisement on the sales of their product. The expenditure on advertisement (in '000 Rs.) and the total profit on sales is given below for a period of 5 years. Find the correlation coefficient and comment on what Raut Pharma ltd. should conclude?

Advertisement expenditure	40	45	48	50	55
Profit in lakhs of Rs.	10	12	11	10	13

20. The following data gives the ages in years of males and their Blood Pressure count. Find the coefficient of correlation.

Age	42	55	63	77	38	49	50	57	60	71
Blood Pressure	122	132	130	143	127	146	152	155	145	158

21. The covariance between two variables is 105 and the standard deviations are 10.5 and 14.2. Find the correlation coefficient.

22. The covariance between two variables is 1065 and the variances are 1210 and 13082. Find the correlation coefficient.

23. If the coefficient of correlation is 0.76 and the standard deviations are 11.54 and 12.8, find the covariance between the two variables.

24. Find the number of observations if r = 0.9, $\sigma_x = 6$, $\sigma_y = 5$ and $\Sigma(x-\bar{x})(y-\bar{y}) = 270$.

25. If for a pair of data, n = 12, Σx = 110, Σy = 90, Σx^2 = 1260, Σy^2 = 950 and Σxy = 1010, find the coefficient of correlation.

26. The following bivariate table gives the ages of couple and the ages of their children. Find the Correlation coefficient.

Ages of parents / Ages of children	20 – 30	30 – 40	40 – 60
0 – 10	18	12	2
10 – 20	4	20	10
20 – 30	-	1	15

27. The following bivariate table gives frequency distribution for the classes. Find the coefficient of correlation.

Class I / Class II	0 – 20	20 – 40	40 – 60	60 – 80
0 – 100	5	11	-	14
100 – 200	10	-	5	10
200 – 300	-	5	15	20
300 – 400	2	6	12	-

28. The following bivariate table gives frequency distribution of sales of two products of a company. Find the coefficient of correlation.

Product I / Product II	5 – 15	15 – 25	25 – 35	35 – 45
30 – 35	-	10	6	4
35 – 40	4	8	-	12
40 – 45	6	-	4	15
300 – 400	10	14	10	-

29. The marks obtained by 20 students in two subjects are given below in pairs. Prepare a bivariate frequency distribution by taking proper class intervals and find the Karl Pearson's coefficient of correlation between the two results:

(10, 12), (7, 9), (8, 16), (12, 6), (5, 9),(4, 2), (8, 10), (11, 16), (13, 7), (9, 12),(17, 19), (15, 19), (10, 11), (6, 8), (3, 10),(5, 5), (2, 9), (9, 4), (7, 2) and (18, 6).

30. The following data gives the ranks given to students in two subjects. Find the rank correlation coefficient.

Sub: A	6	1	3	5	2	4	7	8

Sub: B	3	4	7	2	8	1	5	6

31. The ranks given by judges of a competition to participants are given below. Find the rank correlation coefficient.

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

32. The marks obtained by 10 students in two subjects Business Law (X) and Business Statistics (Y) are given below. Find the rank correlation coefficient.

X	30	78	45	60	59	48	38	77	65	81
Y	42	70	58	65	45	49	40	66	71	63

33. The following data gives the marks obtained by 8 students in their theory and practical examination. Find the rank correlation coefficient.

Theory	45	52	56	41	35	29	53	46
Practical	12	15	16	13	10	8	17	11

34. Find the rank correlation coefficient for the following data:

Height in cm	120	136	178	120	135	160	175	152
Weight in kg	33	41	48	41	50	54	58	41

35. The marks obtained by students in two semesters are given below. Rank the marks and find the rank correlation coefficient.

Sem I	345	440	560	340	345	560	400	380	345	490
Sem II	380	400	520	610	422	400	520	353	518	500

36. Ten candidates, who appeared for a PI, were interviewed by the MD (X), Deputy Manager (Y) and HR Manager (Z). The candidates were given ranks by all the three as shown below. Find using Spearman's Rank Correlation Coefficient, which two of them have a most common approach.

X	5	9	1	6	3	8	4	7	2	10
Y	3	7	2	8	5	6	1	4	10	9

Z	4	8	1	7	6	10	5	3	2	9

37. The salesmen in a company were given marks on the basis of their performance, by the Board of Directors A, B and C. The marks were as follows:

A	11	14	10	18	7	16	9	5	13	20
B	16	18	19	13	20	11	12	6	10	15
C	10	12	11	19	8	17	7	9	14	18

Rank the above data and find the rank correlation between all pairs of Directors, and comment which pair has the most common approach in their assessment.

38. The rank correlation coefficient between two data is given as 0.25. If the sum of the squares of the differences in their ranks is 63, find the number of observations.

39. The rank correlation coefficient between two data is given as 0.5. If the sum of the squares of the differences in their ranks is 110, find the number of observations.

40. If the sum of the square of the difference in the ranks for a bivariate data with repeated ranks is 278, the rank correlation coefficient -0.7 and number of observations 10, find the total correction factor.

Multiple Choice Questions:

1) If r =1, then there is _____ correlation between the two variables.

a) No b) Perfect negative

c) Perfect positive d) Elastic

2) Product moment correlation coefficient is also known as _____

a) Pearson's b) Spearman's

c) Laspeyre's d) Paasche's

3) Coefficient of correlation lies between _____.

a) -1 and +1 b) -2 and +2

c) 0 and -1 d) None of these

4) Find Rank Correlation coefficient if ∑ d^2= 204 and n=10.

a) 0.273 b) -0.237

c) 0.237 d) 0.5

5) When the values of two variables move in the same direction, correlation is said to be .....

a) Linear b) Non-linear c) Positive d) Negative

6) The correlation between shoe-size and intelligence is

a) Zero b) Positive c) Negative

d) None of these

7) Scatter diagram helps us to

a) Find the nature correlation between two variables.

b) Compute the extent of correlation between two variables.

c) Obtain the mathematical relationship between two variables.

d) Both (a) and (c).

8) The covariance between two variables is

a) Strictly positive b) Strictly negative

c) always zero d) either positive or negative or zero

9) For finding correlation between two attributes we consider

a) Pearson's correlation coefficient b) Scatter diagram

c) Spearman's rank correlation coefficient

d) Coefficient of correlation

10) "Demand for goods and their prices under normal times" , correlation are

a) Positive b) negative c) Zero d) None of these

7.11 LIST OF REFERENCES

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

8

REGRESSION ANALYSIS

Unit Structure

8.0 OBJECTIVES:

After going through this chapter you will able to know:

- Meaning of regression.
- Types of methods to solve regression equation.
- Least square method to solve linear regression method.
- Properties of regression analysis.

8.1 INTRODUCTION:

In the previous chapter on Correlation we have seen that the correlation coefficients measure the magnitude and direction of correlation between two variables. The statistical analysts are not satisfied with only the degree of correlation but are also interested to know what the mathematical relation between the variables into consideration is. Obviously, this is only possible when the correlation is due to a cause and effect relation between the concerned variable. If a relation between the expenditure on advertisement and the sales of a product is known to the company, it can easily predict the sales for a particular amount of advertisement. Thus, such a relation is useful in estimation, an important tool of statistical analysis.

For example, if the correlation coefficient between the heights and weights of people is 0.8, it leads to the conclusion that height of a person is strongly and positively related with his or her weight. The obvious interesting question will be, can this relation be represented by a linear (or non- linear) equation. To answer all such questions regression analysis is used.

The study of defining a mathematical relation between two or more variables and facilitate in forecasting or estimating value of one variable given the value of the other variable is called as regression analysis.

Thus, correlation coefficient gives the degree of correlation and regression gives the exact relation between the variables

8.2 IMPORTANCE OF REGRESSION ANALYSIS:

- Regression analysis is statistical technique to represent the relationship between variables. This is used widely in various fields like, social sciences, psychology, economics, bioinformatics, business *etc*.
- The important aspect of regression analysis is its use to estimate value of the dependent variable using the value of the independent variable.
- The mathematical equations representing the relation between variables are called as regression equations and the coefficients of the variables in the equation are called as regression coefficients.
- The correlation coefficient and the regression coefficients are also mathematically related. This helps in estimating the correlation coefficient if the regression coefficients are known.

8.3 METHODS OF STUDYING REGRESSION:

There are two methods to determine the regression equations:

(1) *Free Hand Curve method*: This is a graphical way of drawing regression lines using Scatter Diagram.

(2) *Method of Least Squares*: This is used to determine the regression equations assuming that the relation is linear.

As per our scope of syllabus, we shall study the second method.

8.3.1 Method of Least Squares:

In this method the sum of the squares of the deviations of the values of the variables from its estimated value represented by the best suitable linear equation is minimized.

If $y = a + bx$ is the best fitting line for a given set of data related to variables x and y, then by using the method of least squares we have the conditions as follows:

$$\Sigma y = Na + b\Sigma x \qquad \ldots (1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \ldots (2)$$

These are called as the normal equations. Solving these normal equations simultaneously, we get the values of a and b, the regression coefficients. The equation $y = a + bx$ is called as the *regression equation of y on x.*

If $x = a' + b'y$ is the best fitting line to a given set of data related to variables x and y, then by using the method of least squares we have the conditions as follows:

$$\Sigma x = Na' + b' \Sigma y \qquad \ldots (1)$$

$$\Sigma xy = a'\Sigma y + b'\Sigma y^2 \qquad \ldots (2)$$

These are called as the normal equations. Solving these normal equations simultaneously, we get the values of a' and b', the regression coefficients. The equation $x = a' + b'y$ is called as the *regression equation of x on y*.

1) Direct formula:

For all computational purposes, the following formula is more popular as it uses the direct values of the variables from the raw data.

(*i*) If $y = a + bx$ is a regression equation of y on x, then the regression coefficients a and b are computed as follows:

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \text{ and } a = \frac{\Sigma y}{n} - b_{yx}.\frac{\Sigma x}{n} = \bar{y} - b_{yx}.\bar{x} \qquad \ldots (3)$$

The regression coefficient b is corresponding to the regression equation of y on x, hence is denoted as b_{yx} and is called as the regression coefficient of y on x.

(*ii*) If $x = a' + b'y$ is a regression equation of x on y, then the regression coefficients a' and b' are computed as follows:

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2} \text{ and } a' = \frac{\Sigma x}{n} - b_{yx}\frac{\Sigma y}{n} \qquad \ldots (4)$$

The regression coefficient b' is corresponding to the regression equation of x on y, hence is denoted as b_{xy} and is called as the regression coefficient of x on y.

If we observe the formula for b_{yx} and b_{xy} carefully the denominators are nothing but respective variances (or squares of the respective standard deviations) and the numerator is the same as in the formula for computing the correlation coefficient. This is no coincidence. The regression coefficients are indeed related to the standard deviations and the correlation coefficient. This leads us to the second type of formula as mentioned below

(1) The relation between the regression coefficients, standard deviations and correlation coefficient is as follows:

$$b_{yx} = r\frac{\sigma_y}{\sigma x} \text{ and } b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$\Rightarrow b_{yx} \cdot b_{xy} = \left(r\frac{\sigma_y}{\sigma x}\right)\left(r\frac{\sigma_x}{\sigma_y}\right) = r^2$$

$$\therefore r = \pm\sqrt{b_{yx} \cdot b_{xy}} \qquad \ldots (5)$$

We have mentioned in the section **5.2** that the regression coefficients can be used to compute the correlation coefficient. The above formula no. (5) gives the same.

The sign of *r* depends on the signs of the regression coefficients b_{yx} and b_{xy}.

If both $\boldsymbol{b_{yx}}$ and $\boldsymbol{b_{xy}}$ are positive then ***r*** is also positive.

If both $\boldsymbol{b_{yx}}$ and $\boldsymbol{b_{xy}}$ are negative then ***r*** is also negative.

(2) If the respective means and the regression coefficients are known then the two regression equations are given by the formula:

Regression equation of *y* on *x*: $x - \bar{x} = b_{yx}(y - \bar{y})$ … (6)

Regression Equation of *x* on *y*: $(y - \bar{y}) = b_{xy}(x - \bar{x})$ … (7)

8.4 PROPERTIES OF REGRESSION

(1) The point of intersection of the two regression equations is $(\bar{x}, \bar{y})$. If we solve simultaneously both the regression equations, then the solution set or simply the point of intersection is the mean of *x* and mean of *y*.

(2) The correlation coefficient is the geometric mean of the regression coefficients. *i.e.* $r = \pm\sqrt{b_{yx} \cdot b_{xy}}$. Both the regression coefficients have the same sign (either both are positive or both are negative).

(3) Regression equations are independent of change of origin but not independent of change of scale.

ALL FORMULAE AT A GLANCE

Regression Equation of *y* on *x*	**Regression Equation of *x* on *y***
$y = a + b_{yx}.x$	$x = a' + b_{xy}.y$
$b_{yx} = \dfrac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$ $a = \dfrac{\Sigma y}{n} - b_{yx}.\dfrac{\Sigma x}{n}$	$b_{xy} = \dfrac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$ $a' = \dfrac{\Sigma x}{n} - b_{xy}.\dfrac{\Sigma y}{n}$
Normal Equations:	Normal Equations

$\Sigma y = Na + b\Sigma x$ $\Sigma xy = a\Sigma x + b\Sigma x^2$	$\Sigma x = Na' + b'\Sigma y$ $\Sigma xy = a'\Sigma y + b'\Sigma y^2$
$y - \bar{y} = b_{yx}(x - \bar{x})$	$x - \bar{x} = b_{xy}(y - \bar{y})$
The point of intersection of the regression equations is $(\bar{x}, \bar{y})$	
$r = \pm\sqrt{b_{yx} \cdot b_{xy}}$ where r is positive if both b_{yx} & b_{xy} are positive and r is negative if both b_{yx} & b_{xy} are negative	
The regression coefficients are independent of change of origin.	

Example 1: Find the two regression equations given the following information: $\Sigma xy = 382, \Sigma x = 75$, $\Sigma x^2 = 1442$, $\Sigma y = 70, \Sigma y^2 = 1320$ and $n = 10$.

Solution: (i) Regression Equation of y on x: $y = a + b_{yx}x$

We first find the regression coefficient b_{yx}:

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{10(482) - (75)(70)}{10(1442) - (75)^2} = \frac{4820 - 5250}{14420 - 5625} = -0.049$$

and $a = \frac{\Sigma y}{n} - b_{yx}\frac{\Sigma x}{n} = \frac{70}{10} - (-0.049)\frac{75}{10} = 7 + 0.37 = 7.37$

$\therefore$ Regression Equation of y on x is $y = 7.37 - 0.049x$ … (1)

(ii) Regression Equation of x on y: $x = a' + b_{xy}y$

Now, $b_{xy} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{10(482) - (75)(70)}{10(1320) - (70)^2} = \frac{4820 - 5250}{13200 - 4900} = -0.052$

and $a' = \frac{\Sigma x}{n} - b_{xy}\frac{\Sigma y}{n} = \frac{75}{10} - (-0.052)\frac{70}{10} = 7.5 + 0.364 = 7.864$

$\therefore$ Regression Equation of x on y is $x = 7.864 - 0.052y$ … (2)

Example 2: The following data gives the amount of sales and purchase in lakhs of Rs. of a company. Find the regression equations.

Sales (s)	20	43	22	30	40
Purchase (p)	18	32	15	25	30

Solution: Since the values are large, we shall use the property that the regression coefficients are independent of change of origin.

Let $x = s - 22$ and $y = p - 25$. The table of computations is as shown below:

Sales (s)	$x = s - 22$	*Purchase* (p)	$y = p - 25$	xy	x^2	y^2
20	– 2	18	– 7	14	4	49
43	21	32	7	147	441	49
22	0	15	– 10	0	0	100
30	12	25	0	0	144	0
40	18	30	5	90	324	25
Total	$\Sigma x =$ 43	-	$\Sigma y =$ -5	$\Sigma xy =$ 251	$\Sigma x^2 =$ 913	$\Sigma y^2 =$ 223

(i) To find Regression Equation of y on x: $y = a + b_{yx}x$

Using the table values we have,

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5(251) - (43)(-5)}{5(913) - (43)^2} = \frac{1255 + 215}{4565 - 1849} = 0.54$$

and $a = \frac{\Sigma y}{n} - b_{yx}\frac{\Sigma x}{n} = \frac{-5}{5} - (0.54)\frac{43}{5} = -1 - 4.64 = -5.64$

$\therefore$ Regression Equation of y on x is $y = -5.64 + 0.54x$ … (1)

(ii) Regression Equation of x on y: $x = a' + b_{xy}y$

Now, $b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{5(251) - (43)(-5)}{5(223) - (-5)^2} = \frac{1255 + 215}{1115 - 25} = 1.35$

and $a' = \frac{\Sigma x}{n} - b_{xy}\frac{\Sigma y}{n} = \frac{43}{5} - (1.35)\frac{(-5)}{5} = 8.6 + 1.35 = 9.95$

$\therefore$ Regression Equation of x on y is $x = 9.95 - 1.35y$ … (2)

Example 3: The average marks of 300 students in English and Hindi are 45 and 56 respectively while their respective standard deviations are 10 and 12. If the sum of the products of the deviations from the averages is 32724, find the regression equations. Also estimate the marks obtained in English if a student obtains 70 marks in Hindi.

Solution: Let the data related to English be denoted by x and that related to Hindi be denoted by y. Given: $n = 300$, $\bar{x} = 45$, $\bar{y} = 56$, $\sigma_x = 10$, $\sigma_y = 12$ and $\Sigma(x - \bar{x})(y - \bar{y}) = 32724$.

The problem is solved in 4 steps: (i) finding r, (ii) computing the regression coefficients,

(iii) finding the regression equations and (iv) Estimation.

Now, we know that $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$

$\therefore r = \frac{32724}{300 \text{ x } 10 \text{ x } 12} = 0.909$

The regression coefficients can now be obtained as follows:

$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.909 \text{ x } \frac{12}{10} = 1.09$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.909 \text{ x } \frac{10}{12} = 0.7575$

The regression equation of y on x is given by $y - \bar{y} = b_{yx}(x - \bar{x})$

$\therefore y - 56 = 1.09(x - 45) \quad \Rightarrow y - 56 = 1.09x - 49.05$

$\therefore y = 1.09x - 49.05 + 56$

$\therefore y = 6.95 + 1.09x \quad \ldots (1)$

The regression equation of x on y is given by $x - \bar{x} = b_{xy}(y - \bar{y})$

$\therefore x - 45 = 0.7575(y - 56) \Rightarrow x - 45 = 0.7575y - 42.42$

$\therefore x = 0.7575y - 42.42 + 45$

$\therefore x = 2.58 + 0.7575y \quad \ldots (2)$

Estimation:

To find marks obtained in English (x) if marks obtained in Hindi (y) are 70

Given $y = 70$, to find x we use the regression equation of x on y.

Substituting $y = 70$ in eqn (2) above, we get

$x_{est} = 2.58 + 0.7575(70) = 2.58 + 53.025 = 55.6$

$\therefore x_{est} \approx 56$

Thus, the estimated marks in English are 56.

Example 4: Using the following tabulated information, find (i) the most probable value of x when $y = 10$ and (ii) the most probable value of y when $x = 12$.

$r = 0.65$	x	y
Mean	15	22
S.D.	7.5	9

Solution: Given from the table: $r = 0.65$, $\bar{x} = 15$, $\bar{y} = 22$, $\sigma_x = 7.5$, $\sigma_y = 9$

We first find the regression coefficients:

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.65 \text{ x } \frac{9}{7.5} = 0.78 \text{ and } b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.65 \text{ x } \frac{7.5}{9} = 0.54$$

The regression equation of y on x is given by $y - \bar{y} = b_{yx}(x - \bar{x})$

$\therefore y - 22 = 0.78(x - 15) \quad \Rightarrow y - 22 = 0.78x - 11.7$

$\therefore y = 0.78x - 11.7 + 22$

$\therefore y = 10.3 + 0.78x \quad \ldots (1)$

The regression equation of x on y is given by $x - \bar{x} = b_{xy}(y - \bar{y})$

$\therefore x - 15 = 0.54(y - 22) \Rightarrow x - 15 = 0.54y - 11.88$

$\therefore x = 0.54y - 11.88 + 15$

$\therefore x = 26.88 + 0.54y \quad \ldots (2)$

Estimation:

(i) To estimate the value of x when $y = 10$

Given $y = 10$, to find x we use the regression equation of x on y.

Substituting $y = 10$ in eqn (2) above, we get

$x_{est} = 26.88 + 0.54(10) = 26.88 + 5.4 = 32.28$

(ii) To estimate the value of y when $x = 12$

Given $x = 12$, to find y we use the regression equation of y on x.

Substituting $x = 12$ in eqn (1) above we get

$y_{est} = 10.3 + 0.78(12) = 10.3 + 9.36 = 19.66$

Example 5: The two regression equations are as follows: $4x - 3y + 12 = 0$ and $5x - 2y - 20 = 0$. Find which one of these is the regression equation of y on x and which is the regression equation of x on y.

Solution: Let $4x - 3y + 12 = 0 \quad \ldots (1)$

$5x - 2y - 20 = 0 \quad \ldots (2)$

Since we don't know which equation represents which type, we start with assuming that the eqn (1)is the regression equation of y on x and the eqn (2) is the regression equation of x on y.

Rewriting these equations in their standard form, we have

From (1): $4x - 3y + 12 = 0 \quad \Rightarrow 3y = 4x + 12$

$\therefore y = \frac{4}{3}x + \frac{12}{3} = 4 + 1.33x$

Comparing with $y = a + b_{yx}x$, we have the regression coefficient as $b_{yx} = 1.33$

From (2): $5x - 2y - 20 = 0 \quad \Rightarrow 5x = 2y + 20$

$\therefore x = \frac{2}{5}y + \frac{20}{5} = 4 + 0.4y$

Comparing with $x = a' + b_{xy}y$, we have the regression coefficient as $b_{xy} = 0.4$

Now, we know that $r = \pm\sqrt{b_{yx} \cdot b_{xy}}$

Since both the coefficients are positive, r is also positive and its value is:

$r = \sqrt{1.33 \times 0.4} = \sqrt{0.532} = 0.73$

Since the value of r is in the range of -1 and 1, our assumption about the equations is correct.

Thus eqn(1) represents the regression equation of y on x and eqn(2) represents the regression equation of x on y.

Note:

1. Problems of such type are to be solved by making the assumption as discussed in the previous problem.
2. If the value of r comes out to be greater than 1 or less than -1, then we have to back to the first step and alter the assumption made regarding the equations. Simplify the equations to get the regression coefficients and then find the correct value of r.
3. This method is lengthy only if our assumption is wrong.

Example 6: The following information is provided regarding the regression equation of y on x: The equation is $5x - 2y - 21 = 0$, $\bar{x} = 9$, the coefficient of correlation is 0.8. Find the mean value of y and the ratio of the standard deviations of x and y.

Solution: The given equation $5x - 2y - 21 = 0$ is the regression equation of y on x. Rewriting it in the standard form, we have:

$2y = 5x - 21 \quad \Rightarrow y = 2.5x - 10.5$ (dividing by 2 throughout)… (*)

$\therefore b_{yx} = 2.5$

We know that the point of intersection of the regression equations is the mean value of x and y. In other words $(\bar{x}, \bar{y})$ satisfy the regression equations.

Thus, to find the mean value of y, we substitute $\bar{x} = 9$ in (*)

$\therefore y = 2.5\,(9) - 10.5 = 22.5 - 10.5 = 12$

Mean value of $y = 12$.

Given $r = 0.8$, to find the ratio of the S.D.'s we use the formula: $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$\therefore 2.5 = 0.8 \text{ x } \frac{\sigma_y}{\sigma_x}$

$\therefore \frac{\sigma_y}{\sigma_x} = \frac{2.5}{0.8} = \frac{25}{8}$

$\therefore$ the ratio of standard deviations is 25:8

8.5 LET US SUM UP

In this chapter we have learn:

- Definition of regression and types of regression.
- Using least square methods formed linear regression equation.
- To solve regression equations using analysis
- Using regression equation we can predicate value of unknown variable using known variable.

8.6 UNIT END EXERCISES:

1. Define regression analysis.
2. Explain the significance of regression equations.
3. Define regression coefficients. Also state their properties.
4. Find the two regression equations given the following information: $\Sigma xy = 456, \Sigma x = 90,\ \Sigma x^2 = 920,\ \Sigma y = 80, \Sigma y^2 = 1360$ and $n = 10$.
5. Find the two regression equations given the following information:

   $\Sigma xy = 14345, \Sigma x = 210, \Sigma x^2 = 12342, \Sigma y = 182, \Sigma y^2 = 25720$ and $n = 10$

   Also find the correlation coefficient.
6. Find the two regression equations given the following information:

   $\Sigma xy = 45, \Sigma x = 12, \Sigma x^2 = 32, \Sigma y = 32, \Sigma y^2 = 144$ and $n = 8$

   Also find the correlation coefficient.
7. The following details are given regarding the prices of pulses in Mumbai and Raigad. Average price of pulses in Mumbai: Rs. 14

   S.D. of price of pulses in Mumbai : Rs. 2

   Average price of pulses in Raigad : Rs. 10

   S.D. of price of pulses in Raigad : Rs. 4

If the coefficient of correlation is 0.4, find the (i) price of pulses in Raigad if that in Mumbai are Rs. 10 and (ii) price of pulses in Mumbai, if that in Raigad are Rs. 8.

8. The following data gives the amount of sales in lakhs of Rs. and of two products of a company. Find the regression equations.

Product I	32	78	45	66	89
Product II	21	37	26	33	41

9. By using the normal equations, find the two regression equations. Also estimate the value of (i) y when $x = 18$ and (ii) x when $y = 12$

X	2	5	8	11	14
Y	4	13	22	31	40

10. From the following data find the two regression equations and also the correlation coefficient

X	2	4	1	6	3	8	7	10
Y	8	10	9	11	12	5	6	2

11. The marks obtained by students in two terms, Term I and Term II, are given below. Find the two regression equations and the correlation coefficient. Also estimate (i) marks in Term I if marks in Term II are 40 and (ii) marks in Term II if marks in Term I are 30.

Term I	50	52	47	60	38	79	85	40	55	77
Term II	45	50	49	64	40	80	82	43	51	75

12. Find a regression equation of income (y) on the expenditure (x) on the basis of the following data provided. Estimate the income if the expenditure is Rs.11,550.

Income in '000Rs.	5	12	8	11	16	22	30	25
Expenditure in '000 Rs.	3	7	3	8	11	17	22	20

13. The following data gives the height in cm of 8 mothers and their sons. Find the two regression equations. Estimate the height of a son whose mother's height is 165 cm.

Height of mother	145	166	170	150	155	172	138	159
Height of son	150	162	160	170	154	170	142	161

14. Mr. Sagar Mistry is dealer of four wheeler vehicles and also owns his own garage. The annual maintenance charge is Rs. 600. The following data gives the number of vehicles sold and number of car owners who took AMC from him. Find the two regression equations and estimate the income from AMC's of Mr. Mistry if 300 cars are sold.

No of Cars sold	50	90	65	100	120	180	150	175
No of AMC's	10	36	30	48	54	77	110	125

15. The following data gives the income of 10 persons corresponding to their period of service.

Salary in '000 Rs	4	6	8	14	18	22	28	35	40	45
Period of service	1	3	6	8	10	14	18	20	25	30

Find the two regression equations and estimate (i) period of service, if salary is Rs. 50,000 and (ii) salary, if period of service is 35 years.

16. The following data gives the ages of males (x) and their Blood Pressure (y). Find the regression equations and hence estimate (i) age when blood pressure is 140 and (ii) Blood pressure when age is 40 years.

Age in years	25	28	37	32	45	68	52	78	60	65
Blood Pressure	145	156	160	135	142	155	162	176	166	172

17. Fit a regression line of y on x by method of least square to the following data:

X	2	3	4	5	6	8
Y	2.4	2.9	4.2	5.6	5.5	7.4

18. The average marks of 400 students in Economics and Accounts are 45 and 65 respectively while their respective standard deviations are 12 and 16. If the sum of the products of the deviations from the averages is 32700, find the regression equations. Also estimate the marks obtained in Economics if a student obtains 80 marks in Accounts.

19. The average marks of 100 students in Mathematics and Statistics are 52 and 48 respectively while their respective standard deviations are 9 and 13.5. If the sum of the products of the deviations from the averages is 11450, find the regression equations. Also estimate the marks obtained in Statistics if a student obtains 75 marks in Mathematics.

20. The average marks of 250 students in Business Law and Industrial Law are 40 and 52 respectively while their respective standard deviations are 8 and 11. If the sum of the products of the deviations from the averages is 9875, find the regression equations. Also

estimate the marks obtained in (i) Business Law if a student obtains 60 marks in Industrial Law and (ii) Industrial Law if a student obtains 35 marks in Business Law.

21. Using the following tabulated information, find (i) the most probable value of X when $Y = 10$ and (ii) the most probable value of Y when $X = 15$.

$r = 0.65$	X	Y
Mean	15	22
S.D.	7.5	9

22. Using the following tabulated information, find (i) the most probable value of X when $Y = 25$ and (ii) the most probable value of Y when $X = 30$.

$r = 0.9$	X	Y
Mean	36	38
S.D.	6	8

23. Given the mean wages of men and women working in a factory as Rs. 100 and Rs. 75 with standard deviations Rs. 12 and Rs. 16 and the coefficient of correlation as 0.54, find the regression equations and estimate (i) the wages of men, if the wages for women are Rs. 100 and (ii) the wages of men, if the wages of women are Rs. 60.

24. Find the regression equations, if for a given data $\bar{x} = 120$, $\bar{y} = 150$, $\sigma_x = 16$, $\sigma_y = 12$ and $r = 0.58$.

25. The following data is about the Sales and advertising expenditure of a company A to Z ltd. which produces educational software.

$r = 0.24$	*Sales in '00 Rs.*	*Advertising in '00 Rs.*
Mean	Rs. 45	Rs. 18
S.D.	Rs. 14	Rs. 11

Find the regression equations and estimate the most probable amount of sales if the advertising expenditure is Rs. 25,000.

26. Find the regression equations of y on x and of x on y, if for a given data $\bar{x} = 30$, $\bar{y} = 50$, $\sigma_x = 12.5$, $\sigma_y = 10$ and $r = -0.25$.

27. The following data is given for the marks obtained by students in the subjects of Business Communication and English Literature in an examination. Find the regression equations and estimate (i) the most probable marks in Business Communication if the marks in English Literature are 78.

$r = 0.85$	*Business Communication*	*English Literature*
Mean	64	56

	S.D.	16	12

28. The following equations represent the regression equations. Find the mean values of x an y.

    $3x + 7y = 114$ and $5x + 3y = 86$.

29. The regression equations are givens as: $2x - y = 10$ and $3x - 2y - 5 = 0$. Find the mean values of x and y. Also find the correlation coefficient.

30. The two regression equations are given as : $4x - 3y - 27 = 0$ and $3x - 4y + 6 = 0$. Find

    (i) The mean values of x and y, (ii) correlation coefficient.

31. The two regression equations are as follows: $4x - 3y + 3 = 0$ and $5x - 2y - 26 = 0$. Find (i)which is the regression equation of y on x and which is the regression equation of x on y, (ii)the mean values of x and y, (iii) correlation coefficient, (iv) S.D. of x if S.D. of y is 12 and (v) most probable value of y if $x = 10$.

32. The two regression equations are as follows: $2y - x - 2 = 0$ and $3y - 2x + 1 = 0$. Find (i)which is the regression equation of y on x and which is the regression equation of x on y, (ii)the mean values of x and y, (iii) correlation coefficient, (iii) the most probable value of y when $x = 11$ and (iv) the most probable value of x when $y = 20$.

33. The two regression equations are as follows: $2x - y - 3 = 0$ and $4x - 2y - 6 = 0$. Find (i)which is the regression equation of y on x and which is the regression equation of x on y, (ii)the mean values of x and y, (iii) correlation coefficient, (iii) the most probable value of y when $x = 14$ and (iv) the most probable value of x when $y = 30$.

34. The following information is provided regarding the regression equation of y on x: The equation is $10y - 8x - 240 = 0$, $\bar{y} = 40$, The coefficient of correlation is 0.9. Find the ratio of the standard deviations of x and y.

35. The regression equation of amount of fertilizer (x) used on the yield (y) of the crop is given by $2x - 3y + 2 = 0$. If the average yield of the crop is 12 tons and the ratio of the standard deviations is 9:4, find the coefficient of correlation.

Multiple Choice questions:

1) If the regression equation of X and Y is 5X+7Y=135 , the estimated value of X when Y = 10 is _____

   a) 8 b) 10 c) 5 d) 13

2) If the values of regression coefficient are 0.2 and 0.8 ,then the values of correlation coefficient is _____

   a) 0.6 b) -0.6 c) 0.36 d) 0.4

3) If the values calculated for 10 pairs of x and y variables are: $b_{yx} = \frac{1}{3}$, r = $\frac{1}{2}$, s.d of x=3, then s.d of y is

a) 2 b) 4 c) 1 d) 3

4) The value of correlation coefficient is _____of the two regression coefficients

a) Arithmetic Mean b) Geometric Mean

c) Harmonic Mean d) Standard deviation

5) Identify the triplet of numbers representing values of the two regression confidents and correlation coefficient respectively.

a) 1.5, 1.5, 1.5 b) 1, 1, 1

c) 0.5, -0.5, 0.5 d) 0.4, 1.6, -0.8

6) If the two regression lines are represented by 2x-3y+11=0 and x-2y+9=0, the main value of x and y are _____

a) (7, 5) b) (6, 5)

c) (6, 7) d) (5, 7)

7) For 10 pairs of x and y, variance of x and y are respectively 25 and 49 with means 15 and 8 respectivly. If the covariance between x and y is -24.5, thus regression of coefficient of y on x is _____

a) -0.98 b) -0.5

c) -0.78 d) -0.58

8) If the two regression equation are 40x- 18y=214 and 8x-10y+66=0, then coefficient of correlation is_____

a) 0.65 b) 0.85

c) 0.70 d) 0.60.

9) If regression coefficient of y on x is $\frac{-4}{3}$, coefficient of correlation is -0.8, variance of x is 9, then standard deviation of y is _____

a)3 b) 5

c) 8 d) None of these.

10) Two regression lines corriance with each other , thus the correlation between the variables are _____.

a) perfect positive b) perfect negative

c) both (a) and (b). d) None of these.

8.7 LIST OF REFERENCES

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

9

TIME SERIES

Unit Structure

9.0 OBJECTIVES:

After going through this chapter you will able to know:

- The arrangement data with respect to historical order.
- Study time series data.
- Moving average method to study time series.
- Least square method to study and to predicate the future value of the data.
- Find seasonal indices for quarterly.

9.1 INTRODUCTION:

Swami Vivekananda said once, “Taking the inspiration from the past and keeping an eye on the future if we determine our present, we shall lead to success”. Every business venture needs to know their performance in the past and with the help of some predictions based on that would like to decide their strategy for the present. This is applicable to economic policy makers, meteorological department, social scientists, political analysts etc. Forecasting thus is an important tool in Statistical analysis. Globalization has made the economy more competitive and in turn has made it necessary for all businessmen, entrepreneurs to know about the future of their products, keeping in mind all the constraints.

Forecasting techniques facilitates prediction on the basis of a data available from the past. This data from the past is called as *time series.* A set of observations, of a variable, taken at a regular (fixed and equal) interval of time is called times series. The interval of time may be an hour, a day, a month, a quarter, a year or more than that. Census of population for example, is taken after every ten years. A time series is a bivariate data, with time as the independent variable and the other is the variable under consideration. There are various forecasting methods for time series which enable us to study the variations or trends and estimate the same for the future.

If Y denotes the dependent variable and t denotes the time period, then the relation between both the variables is written as $Y = f(t)$. We say that Y is a function of t and can be represented symbolically as Y_t. If at time period t_1, t_2,…., t_n the values of the dependent variable are Y_1, Y_2,…, Y_n then the pairs (t_1, Y_1), (t_2, Y_2),…., (t_n, Y_n) represent the time series.

9.2 IMPORTANCE OF TIME SERIES ANALYSIS

The analysis of the data in the time series using various forecasting models is called as *time series analysis*. The importance of times series analysis is due to the following reasons:

- *Understanding the past behavior*: This involves identifying the various factors that have affected the variable in the past which will help us for the future prediction.
- *Planning the future actions*: Based on the past patterns or trends and influence of various factors, planning for the future is done using forecasting statistical methods.
- *Comparative study*: Two or more time series data can be compared for studying the impact of various factors affecting the values of the dependent variable. For example the growth of population in two cities can be compared for a fixed period of time or the agricultural production in two areas also can be compared to give an overview of the difference or similarities in the observations and analyse them statistically.

9.3 COMPONENTS OF TIME SERIES

The values of the dependent variable in the times series are affected by many factors which can be seasonal, cyclic or due to some impulsive reasons. The major type of variations, also called as the components of times series are as follows:

1) Secular Trend:

The tendency of a time series of increase, stagnate or decrease over a long period of time is called secular trend or just trend of the time series. Many business and economical activities show a secular trend. For example, the interest rates of banks, the rates in the real estate,

inflation, population, birth-rate, death-rate etc. The term "long period" is a relative term. For time series of inflation it may be a week and for a time series of population the period may be in years. The following two graphs of time series of a population and the time series of the death rate of a city are shown. While the population time series shows an upward trend the death rate series shows a downward (declining) trend.

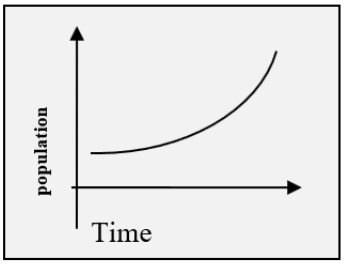


Fig 9.1

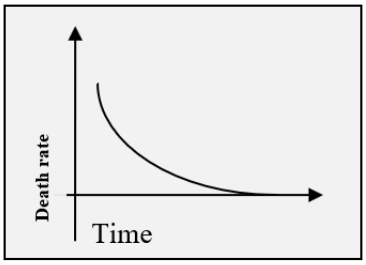


Fig 9.2

The forecasts made on the basis of a secular trend assume that the factors influencing the increase or decrease of the variable are constant throughout a major period of time. However, this may not true always. The secular trend is denoted by *T*.

(1) Short term fluctuations:

(*a*) Seasonal Variations:

The variations that are observed in a time series at a regular interval of time with period less than or equal to a year are called as *seasonal variations.* These occur mainly due to the climatic (seasonal) variations of the nature or due to local traditions and customs. The sales of umbrellas, raincoats in rainy season, woollen clothes in winter season and cotton garments or cold drinks in summer season are example of time series which show seasonal variations due to climatic reasons. The sales and prices of jewellery items in marriage season, crackers in Diwali or sweets during a festival are examples of seasonal variations due to local customs.

The seasonal variations are periodic and regular. The knowledge of such seasonal fluctuations is useful both for the producer from his business point of view and the customer too. The seasonal variations are denoted by *S*.

2) Cyclical Variations:

The variations that are observed in a time series at a regular interval of time with period more than a year are called as *cyclical variations*. Business cycles are examples of such cyclical variations. Here the variations are regular but they may not be periodic. Economic variables show regular ups and downs over a long period of time. These variations may differ in intensity, length of the period. Generally there are four phases of a business cycle as shown below:

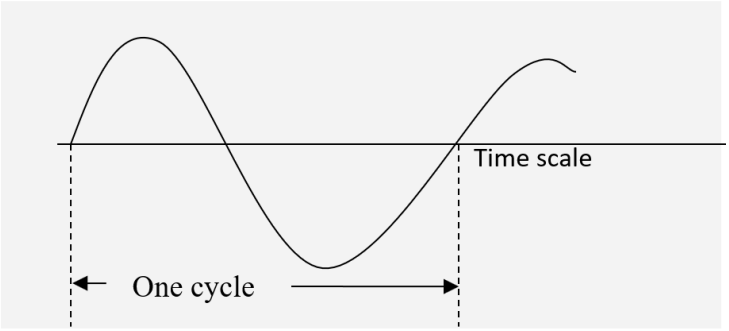


Fig 9.3

(i) The first phase is called as *Boom* which represents prosperity. (ii) Then the next phase is of *Recession* which represents decline. (iii) After recession it is *depression*, which is the lowest peak of the times series. (iv) The final phase is of *recovery*. These phases as mentioned in the beginning are regular but the period of a cycle can vary from two years to ten years or even more in some cases.

The knowledge of cyclic variations is important for a businessman to plan his activity or design his policy for the phase of recession or depression. But one should know that the factors affecting the cyclical variations are quite irregular, difficult to identify and measure. The cyclical variations are denoted by *C*.

3) Irregular Variations:

The variations which occur due random factors are called as *Irregular Variations*. Natural Calamities like earthquake, flood, famines or man made calamities like war, strikes are the major factors responsible for the irregular variations. The occurrence of these variations cannot be predicted and are not bounded by any interval of time. Hence it is difficult to identify and isolate such variations in the time series. The irregular variations are denoted by *I*.

9.4 MODELS FOR ANALYSIS OF TIME SERIES:

As seen above the times series is affected by four components: (i) Secular trend (*T*), (ii) Seasonal variations (*S*), (iii) Cyclic variations (*C*) and (iv) Irregular variations (*I*). The dependent variable can be decomposed into these components using different mathematical models as discussed below:

Additive Model: This model assumes that all the components are independent of each other and their effects on the times series are additive. Symbolically, it can be represented as:

$Y_t = T + S + C + I$

In practice however, the components may be dependent on each other.

Multiplicative Model: This model assumes that the components depend on each other. Symbolically, it can be represented as:

$Y_t = T \times S \times C \times I$

It is also assumed that the geometric mean of the *S*, *C* and *I* is less than or equal to one. *i.e.* $\sqrt{S \times C \times I} \leq 1$.

Though most of the business and economic series follow the multiplicative model, there are mixed models which are also considered wherein one or more of the components are assumed to dependent or independent of the rest. For example:

(i) $Y_t = T \times S \times C + I$ (ii) $Y_t = T + S \times C \times I$ (iii) $Y_t = T + S + C \times I$

All the above mentioned models can be used to determine one unknown value if the remaining values are known.

9.5 METHODS TO FIND TREND:

There are various methods to find the trend. The major methods are as mentioned below:

1. Free Hand Curve.
2. Method of Semi – Averages.
3. Method of Moving Averages.
4. Method of Least Squares.

As per the cope of our syllabus we will restrict our discussion to the method of moving averages and least squares.

9.6 METHOD OF MOVING AVERAGES

This is a simple method in which we take the arithmetic averages of the given times series over a certain period of time. These averages move over that period and are hence called as moving averages. The time interval for the averages is taken as 3 years, 4 years or 5 years and so on. The averages are thus called as 3 yearly, 4 yearly and 5 yearly moving averages. The moving averages are useful in smoothing the fluctuations caused to the variable. Obviously large the time interval of the average more is the smoothing. We shall study the odd yearly (3 and 4) moving averages first and then the 4 yearly moving averages.

Odd Yearly Moving Averages:

In this method the total of the values in the time series is taken for the given time interval and is written in front of the middle value. The average so taken is also written in front of this middle value. This average is the trend for that middle year. The process is continued by replacing the first value with the next value in the time series and so on till the trend for the last middle value is calculated. Let us understand this with examples:

Example 1: Determine the trend for the following data using 3 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997
Sales in '000 Rs.	24	28	30	33	34	36	35	40	44

Solution: The time series is divided into overlapping groups of three years, their 3 yearly totals and averages are calculated as shown in the following table:

Year	Sales (Y)	3 – yearly totals (T)	3 – yearly moving averages: (T/3)
1989	24	-	-
1990	28	24+28+30 = 82	82/3 = 27.33
1991	30	28+30+33 = 91	91/3 = 30.33
1992	33	30+33+34 = 97	97/3 = 32.33
1993	34	33+34+36 = 103	103/3 = 34.33
1994	36	34+36=35 = 105	105/3 = 35
1995	35	36+35+40 = 111	111/3 = 37
1996	40	35+40+44 = 119	119/3 = 39.66
1997	44	-	-

To plot the graph of the actual time series the actual data is plotted and joined by a straight line. Then the trend values (3-yearly moving averages) for the corresponding years are plotted and joined by a dotted line to mark the difference. The graph of the actual time series and the trend values is as shown below:

Example 2: Determine the trend of the following time series using 5 yearly moving averages. Plot the graph of the actual time series and trend values.

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Exports in '000 Rs.	78	84	80	83	86	89	88	90	94	93	96

Solution: The time series is divided into overlapping groups of five years, their 5 yearly totals and averages are calculated as shown in the following table:

The graph of the actual time series and the trend values is as shown below:

Year	Exports (Y)	5 – yearly totals (T)	5 – yearly moving averages: (T/5)
1981	78	-	-
1982	84	-	-
1983	80	78+84+80+83+86 = 411	411/5 = 82.2
1984	83	84+80+83+86+89 = 422	422/5 = 84.4
1985	86	80+83+86+89+88 = 426	85.2
1986	89	83+86+89+88+90 = 436	87.2
1987	88	86+89+88+90+94 = 447	89.4
1988	90	89+88+90+94+93 = 454	90.8
1989	94	88+90+94+93+96 = 461	92.2
1990	93	-	-
1991	96	-	-

Remark:

1. In case of the 3-yearly moving averages, the total and average for the first and the last year in the time series is not calculated. Thus, the moving average of the first and the last year in the series cannot be computed.

2. In case of the 5-yearly moving averages, the total and average for the first two and the last two years in the time series is not calculated. Thus, the moving average of the first two and the last two years in the series cannot be computed.

3. To find the 3-yearly total (or 5-yearly total) for a particular year, you can subtract the first value and add the next value from the previous year's total, so as to save your time!

Even yearly moving averages

In case of even yearly moving averages the method is slightly different as here we cannot find the middle year of the four years in consideration. Here we find the total for the first four years and place it between the second and the third year value of the variable. These totals are again summed into groups of two, called as centered totals and is placed between the two totals. The 4-yearly moving average is found by dividing these centered totals by 8.

Let us understand this method with an example.

Example 3: Calculate the 4 yearly moving averages for the following data:

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Imports in '000 Rs.	15	18	20	24	21	25	28	26	30

Solution: The table of calculation is shown below. Students should leave one line blank after every year to place the centered totals in between two years.

Year	Imports (Y)	4 – yearly totals	4 – yearly centered totals (CT)	4 – yearly moving averages: (CT/8)
1991	15	-	-	-
1992	18	-	-	-
		77		
1993	20		160	20
		83		
1994	24		173	21.6
		90		
1995	21		188	23.5
		98		
1996	25		198	24.8
		100		
1997	28		209	26.1
		109		
1998	26	-	-	-
1999	30	-	-	-

9.7 MERITS AND DEMERITS OF MOVING AVERAGE METHOD

Merits

(1) The method of finding moving averages is very simple and the plotting of trend values is also easy.

(2) There are no mathematical formulae for calculating the trend.

(3) It is more rigidly defined as compared with the free hand curve and semi – averages methods.

(4) In case of cyclical variations, if the period of moving averages and the period of cycle are same then the variations are reduced and in some cases completely eliminated.

(5) It is possible to isolate the fluctuations due to seasonal, cyclical and irregular components using the moving averages method.

Demerits

(1) As mentioned in the remark above, there are no trend values for two or more than two years of the time series.

(2) The method fails in case of non – linear trend.

(3) The trend values obtained from this method do not give any mathematical relationship between time and the variable into consideration. Thus, it is not possible to forecast on the basis of the trend values so obtained.

9.8 METHOD OF LEAST SQUARES:

In this method, which we have already seen in the last chapter, we obtain a straight line trend for the given time series. It is assumed that there is a linear relation between the given bivariate data. This line is called as *the line of best fit.* Let $y = a + bx$ be the line of best fit. We know that the formulae to calculate b and a are as follows:

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \quad \text{and} \quad a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n}$$

Odd number of years in the time series

When the number of years in the given time series is odd, for the middle year we assume the value of $x = 0$. For the years above the middle year the values given to x are …, -2, -1 and so on while those after the middle year are values 1, 2, … and so on.

Even number of years in the time series

When the number of years in the time series is even, then for the upper half years the values of x are assumed as …, -5, -3, -1. For the lower half years, the values of x are assumed as 1, 3, 5, … and so on.

One can easily observe that the values assumed in this fashion will make the total of x to be 0, which simplifies our calculation of the formulae mentioned above. If we substitute $\Sigma x = 0$ in the above formulae, the new formulae for computing the coefficient become as follows:

$$b = \frac{\Sigma xy}{\Sigma x^2} \quad \text{and} \quad a = \frac{\Sigma y}{n}$$

In the table of calculations, now we have to find the values of Σy, Σxy and Σx^2.

Example 4: Fit a straight line trend for the following data giving the annual profits (in lakhs of Rs.) of a company. Estimate the profit for the year 1999.

Year	1992	1993	1994	1995	1996	1997	1998
Profit	30	34	38	36	39	40	44

Solution: Let $y = a + bx$ be the straight line trend.

The number of years is seven, which is odd. Thus, the values of x are taken as 0 for the middle year 1995, for upper three years as -3, -2, -1 and for lower three years as 1, 2, 3.

The table of computation is as shown below:

Year	Profit (y)	X	xy	x^2	Trend Value: $y_t = a + bx$
1992	30	– 3	– 90	9	31.41
1993	34	– 2	– 68	4	33.37
1994	38	– 1	– 38	1	35.33
1995	36	0	0	0	37.29
1996	39	1	39	1	39.25
1997	40	2	80	4	41.21
1998	44	3	132	9	43.17
Total	$\Sigma y = 261$	$\Sigma x = 0$	$\Sigma xy = 55$	$\Sigma x^2 = 28$	

From the table: $n = 7$, $\Sigma xy = 55$, $\Sigma x^2 = 28$, $\Sigma y = 261$

$\therefore b = \frac{\Sigma xy}{\Sigma x^2} = \frac{55}{28} = 1.96$ and $a = \frac{\Sigma y}{n} = \frac{261}{7} = 37.29$

Thus, the straight line trend is $y = 37.29 + 1.96x$.

The trend values in the table for the respective years are calculated by substituting the corresponding value of x in the above trend line equation.

For the trend value for 1992: $x = -3$: $y_{1992} = 37.29 + 1.96 (-3) = 37.29 - 5.88 = 31.41$

Similarly, all the remaining trend values are calculated.

(A short – cut method in case of odd number of years to find the remaining trend values once we calculate the first one, is to add the value of b to the first trend value to get the second trend value, then to the second trend value to get the third one and so on. This is because the difference in the values of x is 1.)

Estimation:

To estimate the profit for the year 1999 in the trend line equation, we substitute the prospective value of x, if the table was extended to 1999. *i.e.* we put $x = 4$, the next value after $x = 3$ for the year 1998.

$\therefore y_{1999} = 37.29 + 1.96 (4) = 45.13$

$\therefore$ the estimated profit for the year 1999 is Rs. 45.13 lakhs.

Example 5: Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2007.

Year	1998	1999	2000	2001	2002	2003	2004	2005
Sales in '000 Rs	120	124	126	130	128	132	138	137

Solution: Let $y = a + bx$ be the straight line trend.

The number of years in the given time series is eight, which is an even number. The upper four years are assigned the values of x as – 7, – 5, – 3, – 1 and the lower four years are assigned the values of x as 1, 3, 5, and 7. Note that here the difference between the values of x is 2, but the sum is zero.

Now, the table of computation is completed as shown below:

Year	Profit (y)	X	xy	x^2	Trend Value: $y_t = a + bx$
1998	120	– 7	– 840	49	120.84
1999	124	– 5	– 620	25	123.28
2000	126	– 3	– 378	9	125.72
2001	130	– 1	– 130	1	128.16
2002	128	1	128	1	130.6
2003	132	3	396	9	133.04
2004	138	5	690	25	135.48
2005	137	7	959	49	137.92
Total	$\Sigma y =$ 1035	Σx =0	$\Sigma xy =$ 205	$\Sigma x^2 =$ 168	

From the table: $n = 8$, $\Sigma xy = 205$, $\Sigma x^2 = 168$, $\Sigma y = 1035$.

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{205}{168} = 1.22 \quad \text{and} \quad a = \frac{\Sigma y}{n} = \frac{1035}{8} = 129.38$$

Thus, the straight line trend is $y = 129.38 + 1.22x$.

The trend values in the table for the respective years are calculated by substituting the corresponding value of x in the above trend line equation.

For the trend value for 1998: $x = – 7$: $y_{1998} = 129.38 + 1.22 (– 7) = 129.38 – 8.54 = 120.84$

Similarly, all the remaining trend values are calculated.

(A short – cut method in case of even number of years to find the remaining trend values once we calculate the first one, is to add twice the value of b to the first trend value to get the second trend value, then to the second trend value to get the third one and so on. This is because the difference in the values of x is 2. In this example we add 2 x 1.22 = 2.44)

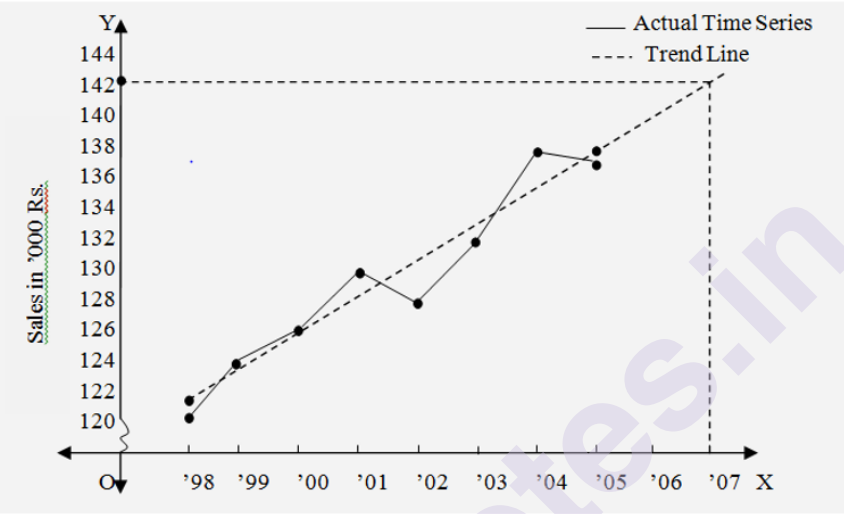
Estimation: To estimate the profit for the year 2007 in the trend line equation, we substitute the prospective value of x, if the table was extended to 2007. *i.e.* we put $x = 11$, the next value after $x = 9$ for the year 2006 and $x = 7$ for 2005.

$\therefore y_{2007} = 129.38 + 1.22 (11) = 142.8$

$\therefore$ the estimated sales for the year 2007 is Rs. 1,42,800.

Now we draw the graph of actual time series by plotting the sales against the corresponding year. The period is taken on the X-axis and the sales on the Y-axis. The points are joined by straight lines. To draw the trend line it is enough to plot any two points (usually we take the first and the last trend value) and join it by straight line.

To estimate the trend value for the year 2007, we draw a line parallel to Y-axis from the period 2007 till it meets the trend line at a point say *A*. From this point we draw a line parallel to the X-axis till it meets the Y-axis at point say *B*. This point is our estimated value of sales for the year 2007. The graph and its estimated value (graphically) is shown below:



From the graph, the estimated value of the sales for the year 2007 is 142 *i.e.* Rs. 1,42,000 (approximately)

Example 6: Fit a straight line trend to the following data. Estimate the import for the year 1998.

Year	1991	1992	1993	1994	1995	1996
Imports in '000 Rs	40	44	48	50	46	52

Solution: Here again the period of years is 6 *i.e.* even. Proceeding similarly as in the above problem, the table of calculations and the estimation is as follows:

Year	Import (y)	X	xy	x^2	Trend Value: $y_t = a + bx$
1991	40	– 5	– 200	25	41.82
1992	44	– 3	– 132	9	43.76
1993	48	– 1	– 48	1	45.7
1994	50	1	50	1	47.64
1995	46	3	138	9	49.58
1996	52	5	260	25	51.52
Total	$\Sigma y = 280$	$\Sigma x = 0$	$\Sigma xy = 68$	$\Sigma x^2 = 70$	

From the table: $n = 6$, $\Sigma xy = 68$, $\Sigma x^2 = 70$, $\Sigma y = 280$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{68}{70} = 0.97 \quad \text{and} \quad a = \frac{\Sigma y}{n} = \frac{280}{6} = 46.67$$

Thus, the straight line trend is $y = 46.67 + 0.97x$.

All the remaining trend values are calculated as described in the above problem.

Estimation: To estimate the imports for the year 1998, we put $x = 9$ in the trend line equation.

$\therefore y_{1997} = 46.67 + 0.97\ (9) = 55.4$

$\therefore$ the imports for the year 1997 are Rs. 55, 400.

9.9 MEASUREMENT OF SEASONAL TREND USING SIMPLE AVERAGE

Seasonal variations as we know have a period of less than a year. All the trend values we have calculated so far have the original time series data with a period difference of one year. Thus, it is not possible to identify and eliminate the seasonal variations using these methods. This process of elimination is called as *deseasonalization*. One of the simple methods used for this is the *simple average method*.

Example 7: Compute the seasonal indices for the following data giving the amount of loan disbursements (in crores of Rs.) of a bank for the years 2004, 2005 and 2006.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2004	22	25	30	35	38	32	30	28	30	36	37	32
2005	23	25	27	33	39	34	29	26	25	34	38	30
2006	21	22	27	32	37	35	28	27	27	32	40	33

Solution: In this case, when the time series data is given month wise we proceed as follows:

The averages of monthly values for all the three years are calculated. For example, the values for the month of January for the three years are: 22, 23 and 21. The average of these values is 22. In a similar way we find all averages of monthly values in the column heading *SA* which stands for seasonal average.

Then we sum all the averages of monthly values *i.e.* we find $\Sigma(SA)$.

Now, the grand average (G) is calculated *i.e.* $G = \frac{\Sigma(SA)}{12}$.

The seasonal index (SI) is now calculated by the formula: $SI = \frac{SA}{G} \times 100$.

For example, the SA value for January is 22, thus its $SI = \frac{22}{30.75} \times 100 = 71.55$

Month	Year			Monthly Totals	*SA*	*SI*
	2004	2005	2006			
Jan	22	23	21	66	22	71.55
Feb	25	25	22	72	24	78.05
Mar	30	27	27	84	28	91.06
Apr	35	33	32	100	33.33	108.39
May	38	39	37	114	38	123.42
Jun	32	42	35	109	36.33	118.15
Jul	30	29	28	87	29	94.31
Aug	28	26	27	81	27	87.80
Sep	30	25	27	82	27.33	88.88
Oct	36	34	32	102	34	110.57
Nov	37	38	40	115	38.33	124.65
Dec	32	30	33	95	31.66	102.96
				$\Sigma(SA) = 368.98$		1199.79
				$G = 30.75$		

The total of all *SI*'s should be 1200 for this example, but we can see that it is 1199.79. This can be adjusted by multiplying each *SI* by the factor: $\frac{1200}{1199.79}$.

Example 8: Compute the seasonal indices for the following data using simple average method:

Quarters	I	II	III	IV
1999	30	34	42	38
2000	32	38	40	42
2001	34	40	44	48
2002	30	36	40	44

Solution: The procedure is analogous to that of the previous one.

The quarterly totals are calculated and the seasonal average for the four years is computed.

The grand average is calculated using the formula: $G = \frac{\Sigma(SA)}{4}$

The Seasonal Index for every quarter is calculated by the formula: $SI = \frac{SA}{G}$ x 100

The table of calculations is as follows:

Quarters	I	II	III	IV
1999	30	34	42	38
2000	32	38	40	42
2001	34	40	44	48
2002	30	36	40	44

Total	126	148	166	172		
SA	31.5	37	41.5	43	$\Sigma(SA)$ =153	G = 38.25
SI	82.35	96.73	108.5	112.42	400	

To find the deseasonalized values of the variable are calculated by the simple formula as given below:

$$\text{Deseasonalized value} = \frac{\text{actual value}}{\text{corresponding } SI} \times 100.$$

Let us calculate the deseasonalized values for the problem.

For example 8, the deseasonalized values are as shown below:

Quarters	I	II	III	IV
1999	36.43	35.15	38.71	33.8
2000	38.86	39.28	36.87	37.36
2001	41.23	41.35	40.55	42.7
2002	36.43	37.22	36.87	39.14

Despite being the simplest method, the method of simple averages is of very little use. The main reason is that it assumes that there is negligible trend or cylical or irregular component in the given time series, which is not practically true. Hence it is not useful for forecasting in business and economical time series data.

9.10 LET US SUM UP

In this chapter we have learn:

- That data can be easy to understand with respect to time.
- To study time series with respect to trend value.
- Moving average method to calculate trend value.
- Least square methods to find trend value and predicate the future value.
- To calculate Seasonal indices.

9.11 UNIT END EXERCISES

1. Define time series and time series analysis.
2. What are the objectives of a time series analysis?
3. Write a short note on short term fluctuations in a time series.
4. What are the different components of a time series? Explain with suitable examples.
5. Write a short note on the different phases of a business cycle.
6. Give the merits and demerits of moving average method.

7. "Longer the period of moving average, better is the trend value obtained", is this statement correct? Justify your answer.
8. Write a short not on the method of least squares.
9. Explain the simple average method to find the seasonal indices of a time series.
10. Define deseasonalization. What is its significance?
11. Determine the trend for the following data using 3 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1977	1978	1979	1980	1981	1982	1983	1984
Sales in lakhs of Rs.	46	54	52	56	58	62	59	63

12. Determine the trend for the following data using 3 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1979	1980	1981	1982	1983	1984	1985	1986
Profit in lakhs of Rs.	98	100	97	101	107	110	102	105

13. Determine the trend for the following data using 5 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1980	1082	1984	1986	1988	1990	1992	1994	1996	1998	2000
Values	34	37	35	38	37	40	43	42	48	50	52

14. Determine the trend for the following data giving the production of steel in million tons, using 5 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Production	28	30.5	32	36.8	38	36	39.4	40.6	42	45	43.5

15. Determine the trend for the following data giving the production of wheat in thousand tons from the year 1980 to 1990, using the 5 – yearly moving averages. Plot the graph of the actual time series and the trend values.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production	13.5	14.7	17	16.2	18.1	20.4	22	21.2	24	25	26.6

16. Determine the trend for the following data giving the income (in million dollars) from the export of a product from the year 1988 to 1999. Use the 4 – yearly moving averages method and plot the graph of actual time series and trend values.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Income	340	360	385	470	430	444	452	473	490	534	541	576

17. Using the 4 – yearly moving average method find the trend for the following data:

Year	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
Value	102	100	103	105	104	109	112	115	113	119	117

18. Determine the trend for the following data giving the sales (in '00 Rs.) of a product per week for 20 weeks. Use appropriate moving average method.

Week	1	2	3	4	5	6	7	8	9	10
Sales	22	26	28	25	30	35	39	36	30	32
Week	11	12	13	14	15	16	17	18	19	20
Sales	29	34	36	35	35	39	43	48	52	49

19. An online marketing company works 5-days a week. The day-to-day total sales (in '000 Rs.) of their products for 4 weeks are given below. Using a proper moving average method find the trend values.

Day	1	2	3	4	5	6	7	8	9	10
Sales	12	16	20	17	18	20	26	25	27	30
Day	11	12	13	14	15	16	17	18	19	20
Sales	35	32	32	38	36	35	34	38	40	41

20. Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2001.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Profit in '000 Rs	76	79	82	84	81	84	89	92	88	90

21. Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2007.

Year	1998	1999	2001	2002	2003	2004	2005	2006
Profit in '000 Rs	116	124	143	135	138	146	142	152

22. Fit a straight line trend for the following data giving the number of casualties (in hundreds) of motorcyclists without helmets. Estimate the number for the year 1999.

Year	1992	1993	1994	1995	1996	1997	1998
No of casualties	12	14.2	15.3	16	18.8	19.6	22.1

23. Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2002.

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Imports in '000 Rs	55	52	50	53	54	56	58	60	57	59

24. Fit a straight line trend to the following data giving the price of crude oil per barrel in USD. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2003.

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Price per barrel	98	102	104.5	108	105	109	112	118	115	120

25. The following data shows the quarterly profits (in crores of Rs.) of an IT company for the years 2002 to 2004. Assuming that the trend is negligible, calculate the seasonal indices using simple average method.

Quarters	I	II	III	IV
2002	42	46	49	52
2003	50	52	55	48
2004	48	50	53	50

26. Assuming that the trend is absent, find the seasonal indices for the following data and also find the deseasonalized values.

Quarters	I	II	III	IV
1977	10	12	14	16
1978	12	15	18	22
1979	16	18	20	24
1980	24	26	28	34

27. Assuming that the trend is absent, find the seasonal indices for the following data and also find the deseasonalized values.

Quarters	I	II	III	IV
1989	21	23	22	24
1990	23	26	27	29
1991	27	30	28	30
1992	29	31	30	33

Multiple choice questions:

1) A time series consists of:
   a) Short-term variations b) Long-term variations
   c) Irregular variations d) All of the above
2) The method of moving average is used to find the:
   a) Secular trend b) Seasonal variation
   c) Cyclical variation d) Irregular variation.
3) Value of b in the trend line Y = a + bX is:
   a) Always negative b) Always positive
   c) Always zero d) Both negative and positive
4) A time series has:
   a) Two components b) Three components
   c) Four components d) Five components
5) For odd number of year, formula to code the values of X by taking origin at Centre is:
   a) X = year – average of years b) X = year – first year
   c) X = year – last year d) X = year – ½ average of years

9.12 LIST OF REFERENCES

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

10

INDEX NUMBER

Unit Structure

10.0 OBJECTIVES:

After going through this chapter you will able to know:

- Definition of index number and its important.
- Types of index numbers.
- The different methods of construction of index number.
- Different applications of index number.

10.1 INTRODUCTION:

Every variable undergoes some changes over a period of time or in different regions or due to some factors affecting it. These changes are needed to be measure. In the last chapter we have seen how a time series helps in estimating the value of a variable in future. But the magnitude of the changes or variations of a variable, if known, are useful for many more reasons. For example, if the changes in prices of various household commodities are known, one can plan for a proper budget for them in

advance. If a share broker is aware of the magnitude of fluctuations in the price of a particular share or about the trend of the market he can plan his course of action of buying or selling his shares. Thus, we can feel that there is a need of such a measure to describe the changes in prices, sales, profits, imports, exports etc, which are useful for a common man to a business organization.

Index number is an important statistical relative tool to measure the changes in a variable or group of variables with respect to time, geographical conditions and other characteristics of the variable(s). Index number is a relative measure, as it is independent of the units of the variable(s) taken in to consideration. This is the advantage of index numbers over normal averages. All the averages which we studied before are absolute measures, *i.e.* they are expressed in units, while index numbers are percentage values which are independent of the units of the variable(s). In calculating an index number, a base period is considered for comparison and the changes in a variable are measured using various methods.

Though index numbers were initially used for measuring the changes in prices of certain variables, now it is used in almost every field of physical sciences, social sciences, government departments, economic bodies and business organizations. The gross national product (GNP), per capita income, cost of living index, production index, consumption, profit/loss etc every variable in economics uses this as a tool to measure the variations. Thus, the fluctuations, small or big, in the economy are measure by index numbers. Hence it is called as a barometer of economics.

10.2 IMPORTANCE OF INDEX NUMBERS

The important characteristics of Index numbers are as follows:

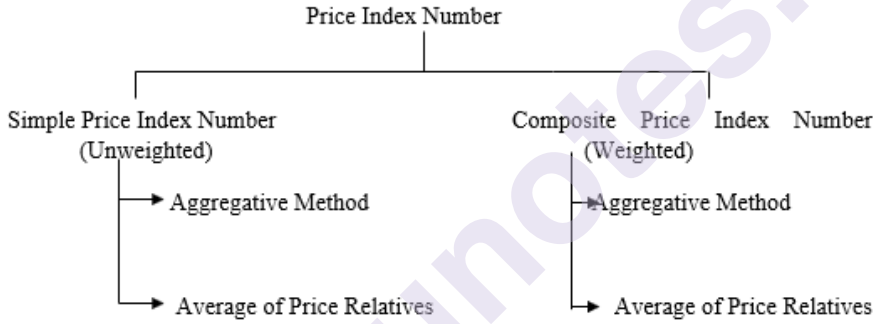
1. *It is a relative measure*: As discussed earlier index numbers are independent of the units of the variable(s), hence it a special kind of average which can be used to compare different types of data expressed in different units at different points of time.
2. *Economical Barometer*: A barometer is an instrument which measure the atmospheric pressure. As the index numbers measure all the ups and downs in the economy they are hence called as the economic barometers.
3. *To generalize the characteristics of a group*: Many a times it is difficult to measure the changes in a variable in complete sense. For example, it is not possible to directly measure the changes in a business activity in a country. But instead if we measure the changes in the factors affecting the business activity, we can generalize it to the complete activity. Similarly the industrial production or the agricultural output cannot be measured directly.
4. *To forecast trends*: Index numbers prove to be very useful in identifying trends in a variable over a period of time and hence are used to forecast the future trends.

5. *To facilitate decision making*: Future estimations are always used for long term and short term planning and formulating a policy for the future by government and private organizations. Price Index numbers provide the requisite for such policy decisions in economics.
6. *To measure the purchasing power of money and useful in deflating*: Index numbers help in deciding the actual purchasing power of money. We often hear from our elders saying that " In our times the salary was just Rs. 100 a month and you are paid Rs. 10,000, still you are not happy!" The answer is simple (because of index numbers!) that the money value of Rs. 100, 30 years before and now is drastically different. Calculation of *real income* using index numbers is an important tool to measure the actual income of an individual. This is called as deflation.

There are different types of index numbers based on their requirement like, price index, quantity index, value index etc. The price index is again classified as single price index and composite price index.

10.3 price index numbers

The price index numbers are classified as shown in the following diagram:



Notations:

P_0: Price in Base Year **Q_0: Quantity in Base Year**

P_1: Price in Current Year **Q_1: Quantity in Current Year**

The suffix '0' stands for the base year and the suffix '1' stands for the current year.

10.3.1 Simple (unweighted) price index no. by aggregative method:

In this method we define the price index number as the ratio of sum of prices in current year to sum of prices in base year and express it in percentage. *i.e.* multiply the quotient by 100.

Symbolically, $I = \frac{\Sigma P_1}{\Sigma P_0} \times 100.$ … (1)

Steps for computation:

1. The total of all base year prices is calculated and denoted by ΣP_0.
2. The total of all current year prices is calculated and denoted by ΣP_1.

3. Using the above formula, simple price index number is computed.

Example 1: From the following data, construct the price index number by simple aggregative method:

Commodity	Unit	Price in 1985	Price in 1986
A	Kg	10	12
B	Kg	4	7
C	Litre	6	7
D	Litre	8	10

Solution: The totals of the 3rd and 4th columns are computed as shown below:

Commodity	Unit	Price in 1985(P_0)	Price in 1986(P_1)
A	Kg	10	12
B	Kg	4	7
C	Litre	6	7
D	Litre	8	10
Total	--	$\Sigma P_0 = 28$	$\Sigma P_1 = 36$

$$\therefore I = \frac{\Sigma P_1}{\Sigma P_0} \text{ x } 100 = \frac{36}{28} \text{ x } 100 = 128.57$$

Meaning of the value of I: I = 128.57 means that the prices in 1986, as compared with that in 1985 have increased by 28.57 %.

10.3.2 Simple (unweighted) price index number by average of price relatives method

In this method the price index is calculated for every commodity and its arithmetic mean is taken. *i.e.* the sum of all price relative is divided by the total number of commodities.

Symbolically, if there are *n* commodities in to consideration, then the simple price index number of the group is calculated by the formula:

$$I = \frac{1}{n}\Sigma\left(\frac{P_1}{P_0} \text{ x } 100\right) \qquad \ldots (2)$$

Steps for computation

1. The price relatives for each commodity are calculated by the formula: $\frac{P_1}{P_0} \text{ x } 100$.

2. The total of these price relatives is calculated and denoted as $\Sigma\left(\frac{P_1}{P_0} \text{ x } 100\right)$.

3. The arithmetic mean of the price realtives using the above formula no. (2) gives the required price index number.

Example 2: Construct the simple price index number for the following data using average of price relatives method:

Commodity	Unit	Price in	
		1997	1998
Rice	Kg	10	13
Wheat	Kg	6	8
Milk	Litre	8	10
Oil	Litre	15	18

Solution: In this method we have to find price relatives for every commodity and then total these price relatives. Introducing the column of price relatives the table of computation is as follows:

Commodity	Unit	Price in		$\frac{P_1}{P_0}$ x 100
		1997(P_0)	1998(P_1)	
Rice	Kg	10	13	130
Wheat	Kg	6	8	133.33
Milk	Litre	8	10	125
Oil	Litre	15	18	120
				Total: 508.33

Now, $n = 4$ and the total of price relatives is 508.33

$$\therefore I = \frac{1}{n}\Sigma\left(\frac{P_1}{P_0} \text{ x } 100\right) = \frac{508.33}{4} = 127.08$$

The prices in 1998 have increased by 27 % as compared with in 1997.

Remark:

1. The simple aggregative method is calculated without taking in to consideration the units of individual items in the group. This may give a misleading index number.

2. This problem is overcome in the average of price relatives method as the individual price relatives are computed first and then their average is taken.

3. Both the methods are unreliable as they give equal weightage to all items in consideration which is not true practically.

10.3.3 Weighted index numbers by aggregative method:

In this method weights assigned to various items are considered in the calculations. The products of the prices with the corresponding weights are computed; their totals are divided and expressed in percentages.

Symbolically, if W denotes the weights assigned and P_0, P_1 have their usual meaning, then the weighted index number using aggregative method is given by the formula:

$$I = \frac{\Sigma P_1 W}{\Sigma P_0 W} \text{ x } 100 \qquad \ldots (3)$$

Steps to find weighted index number using aggregative method

1. The columns of P_1W and P_0W are introduced.
2. The totals of these columns are computed.
3. The formula no. (3) is used for computing the required index number.

Example 3: From the following data, construct the weighted price index number:

Commodity	A	B	C	D
Price in 1982	6	10	4	18
Price in 1983	9	18	6	26
Weight	35	30	20	15

Solution: Following the steps mentioned above the table of computations is as shown below:

Commodity	Weight (W)	Price in 1982 (P_0)	P_0W	Price in 1983 (P_1)	P_1W
A	35	6	210	9	315
B	30	10	300	18	540
C	20	4	80	6	120
D	15	18	270	26	390
Total	-	-	ΣP_0W = 860	-	ΣP_1W = 1365

Using the totals from the table, we have

$$\text{Weighted Index Number } I = \frac{\Sigma P_1 W}{\Sigma P_0 W} \text{x} 100 = \frac{1365}{860} \text{ x } 100 = 158.72$$

Remark: There are different formulae based on what to be taken as the weight while calculating the weighted index numbers. Based on the choice of the weight we are going to study here four types of weighted index numbers: (1) Laspeyre's Index Number, (2) Paasche's Index Number, (3) Fisher's Index Number and (4) Kelly's Index Numbers.

(1) Laspeyre's Index Number:

In this method Laspeyre assumed the base quantity (Q_0) as the weight in constructing the index number. Symbolically, P_0, P_1 and Q_0 having their usual meaning, the Laspeyre's index number denoted by I_L is given by the formula: $I_L = \frac{\Sigma P_1 Q_0}{\Sigma P_0 Q_0} \times 100$ … (4)

Steps to compute I_L:

1. The columns of the products P_0Q_0 and P_1Q_0 are introduced.

2. The totals of these columns are computed.

3. Using the above formula no. (4), I_L is computed.

Example 4: From the data given below, construct the Laspeyre's index number:

Commodity	1965		1966
	Price	Quantity	Price
A	5	12	7
B	7	12	9
C	10	15	15
D	18	5	20

Solution: Introducing the columns of the products P_0Q_0 and P_1Q_0, the table of computation is completed as shown below:

Commodity	1965		1966		
	Price (P_0)	Quantity(Q_0)	Price (P_1)	P_0Q_0	P_1Q_0
A	5	12	7	60	84
B	7	12	9	84	108
C	10	15	15	150	225
D	18	5	20	90	100
Total	-	-	-	ΣP_0Q_0 = 384	ΣP_1Q_0 = 517

Using the totals from the table and substituting in the formula no. (4), we have

$$I_L = \frac{\Sigma P_1 Q_0}{\Sigma P_0 Q_0} \text{ x } 100 = \frac{517}{384} \text{ x } 100 = 134.64$$

(2) Paasche's Index Number:

In this method, Paasch assumed the current year quantity (Q_1) as the weight for constructing the index number. Symbolically, P_0, P_1 and Q_1 having their usual meaning, the Paasche's index number denoted by I_P is given by the formula: $I_P = \frac{\Sigma P_1 Q_1}{\Sigma P_0 Q_1} \text{ x } 100$ … (5)

The steps for computing I_P are similar to that of I_L.

Example 5: From the data given below, construct the Paasche's index number:

Commodity	1985	1986	
	Price	Price	Quantity
A	5	8	10
B	10	14	20
C	6	9	25
D	8	10	10

Solution: Introducing the columns of the products P_0Q_1 and P_1Q_1, the table of computations is completed as shown below:

Commodity	1985	1986			
	Price (P_0)	Price (P_1)	Quantity (Q_1)	P_0Q_1	P_1Q_1
A	5	8	10	50	80
B	10	14	20	200	280
C	6	9	25	150	225
D	8	10	10	80	100
Total	-	-	-	$\Sigma P_0Q_1 = 480$	$\Sigma P_1Q_1 = 685$

Using the totals from the table and substituting in the formula no. (5), we have

$$I_P = \frac{\Sigma P_1 Q_1}{\Sigma P_0 Q_1} \text{ x } 100 = \frac{685}{480} \text{ x } 100 = 142.71$$

(3) Fisher's Index Number:

Fisher developed his own method by using the formulae of Laspeyre and Paasche. He defined the index number as the geometric mean of I_L and I_P. Symbolically, the Fisher's Index number denoted as I_F is given by the formula: $I_F = \sqrt{I_L \text{ x } I_P} = \sqrt{\frac{\Sigma P_1 Q_0}{\Sigma P_0 Q_0} \text{ x } \frac{\Sigma P_1 Q_1}{\Sigma P_0 Q_1}}$ x 100.. (6)

Note:

1. *The multiple 100 is outside the square root sign.*
2. *While computing products of the terms, care should be taken to multiply corresponding numbers properly.*

Example 6: From the following data given below, construct the (i) Laspeyre's index number, (ii) Paasche's index number and hence (iii) Fisher's index number.

Item	1975		1976	
	Price	Quantity	Price	Quantity
A	4	12	6	16
B	2	16	3	20
C	8	9	11	14

Solution: Introducing four columns of the products of P_0Q_0, P_0Q_1, P_1Q_0 and P_1Q_1, the table of computations is completes as shown below:

Item	P_0	Q_0	P_1	Q_1	P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1
A	4	12	6	16	48	64	72	96
B	2	16	3	20	32	40	48	60
C	8	9	11	14	72	112	99	154
				Total	152	216	219	310

From the table, we have $\Sigma P_0Q_0 = 152$, $\Sigma P_0Q_1 = 216$, $\Sigma P_1Q_0 = 219$ and $\Sigma P_1Q_1 = 310$

$$\therefore I_L = \frac{\Sigma P_1 Q_0}{\Sigma P_0 Q_0} \text{x } 100 = \frac{219}{152} \text{x } 100 = 144.08$$

$$\therefore I_P = \frac{\Sigma P_1 Q_1}{\Sigma P_0 Q_1} \text{x } 100 = \frac{310}{216} \text{x } 100 = 143.52$$

$$\therefore I_F = \sqrt{I_L \text{ x } I_P} = \sqrt{144.08 \text{ x } 143.52} = 143.8$$

Remark:

1. Laspeyre's index number though popular has a drawback that it does not consider the change in consumption over a period. (as it does not take into account the current quantity).

2. Paasche's index number overcomes this by assigning the current year quantity as weight.

3. Fisher's index number being the geometric mean of both these index numbers, it considers both the quantities. Hence it is called as the ideal index number.

(4) Kelly's Index Number:

In this method, Kelly assigns the average of both the quantities as the weight for constructing the index number. Symbolically, P_0, P_1, Q_0 and Q_1 having their usual meaning, the Kelly's index number denoted by I_K is given by the formula:

$$I_K = \frac{\Sigma P_1 Q}{\Sigma P_0 Q} \text{ x } 100, \text{ where } Q = \frac{Q_0 + Q_1}{2} \qquad \ldots (7)$$

Steps to find I_K:

1. The average of both the quantities is computed.
2. The column products of the type P_1Q and P_0Q are introduced.
3. The corresponding column totals are computed.
4. Using the formula no. (7), the required index number I_K is computed.

Example 7: From the following data given below, construct the Kelly's index number:

Item	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	18	20	24	22
B	9	10	13	16
C	10	15	12	19
D	6	13	8	15
E	32	14	38	18

Solution: Introducing the columns of $Q = \frac{Q_0 + Q_1}{2}$, P_0Q and P_1Q, the table of computations is completed as shown blow:

Item	Q_0	Q_1	Q	P_0	P_0Q	P_1	P_1Q
A	20	22	21	18	378	24	504
B	10	16	13	9	117	13	169
C	15	19	17	10	170	12	204

D	13	15	14	6	84	8	112
E	14	18	16	32	512	38	608
					1261		1597

From the table, we have $\Sigma P_0Q = 1261$ and $\Sigma P_1Q = 1597$

$$\therefore I_K = \frac{\Sigma P_1Q}{\Sigma P_0Q} \text{ x } 100 = \frac{1597}{1261} \text{ x } 100 = 126.65$$

10.3.4 Weighted index numbers using average of price relatives method:

This is similar to what we have seen in subsection 7.3.2. Here the individual price relatives are computed first. These are multiplied with the corresponding weights. The ratio of the sum of the products and the total value of the weight is defined to be the weighted index number.

Symbolically, if W denotes the weights and I denote the price relatives then the weighted index number is given by the formula: $\frac{\Sigma IW}{\Sigma W}$ … (8)

One of the important weighted index number is the cost of living index number, also known as the consumer price index (CPI) number.

10.4 COST OF LIVING INDEX NUMBER OR CONSUMER PRICE INDEX NUMBER:

There are two methods for constructing this index number: (1) Aggregative expenditure method and (2) Family Budget Method

(1) In aggregative expenditure method we construct the index number by taking the base year quantity as the weight. In fact this index number is nothing but the Laspeyre's index number.

(2) In family budget method, value weights are computed for each item in the group and the index number is computed using the formula:

$$\frac{\Sigma IW}{\Sigma W}, \text{ where } I = \frac{P_1}{P_0} \text{ x } 100 \text{ and } W = P_0Q_0 \quad \text{… (9)}$$

Example 8: A survey of families in a city revealed the following information:

Item	Food	Clothing	Fuel	House Rent	Misc.
% Expenditure	30	20	15	20	15
Price in 1987	320	140	100	250	300

Price in 1988	400	150	125	250	320

What is the cost of living index number for 1988 as compared to that of 1987?

Solution: Here % expenditure is taken as the weight (W). The table of computations are completed as shown below:

Item	P_0	P_1	$I = \frac{P_1}{P_0} \times 100$	% Expenditure (W)	IW
Food	320	400	125	30	3750
Clothing	140	150	107.14	20	2142.8
Fuel	100	125	125	15	1875
House Rent	250	250	100	20	2000
Miscellaneous	300	320	106.67	15	1600.05
Total				$\Sigma W = 100$	11367.85

From the table, we have $\Sigma W = 100$ and $\Sigma IW = 11376.85$

$$\therefore \text{cost of living index number} = \frac{\Sigma IW}{\Sigma W} = \frac{11367.85}{100} = 113.68$$

Use of Cost of living index numbers

1. These index numbers reflect the effect of rise and fall in the economy or change in prices over the standard of living of the people.
2. These index numbers help in determining the purchasing power of money which is the reciprocal of the cost of living index number.
3. It is used in deflation. *i.e.* determining the actual income of an individual. Hence it also used by the management of government or private organizations to formulate their policies regarding the wages, allowance to their employees.

10.5 CHAIN BASE INDEX NUMBER:

In constructing an index number, the reference year called as the base year may be fixed or changing. The indices which are computed using a fixed base are called as **F**ixed **B**ase **I**ndices (FBI) and the indices which are compute using a changing base are called as **C**hain **B**ase **I**ndices (CBI). The CBI's are computed by taking every time the index of the previous year as

the base year. These are called as *link relatives*. The chain index is now calculated by the formula:

$$\text{Chain Index for a year} = \frac{\text{link relative of the year x chain index of previous year}}{100} \quad \ldots \quad (10)$$

Conversion From Fixed Base To Chain Base

A FBI is converted to a CBI by the following formula:

$$\text{CBI of a year} = \frac{\text{FBI of the year x 100}}{\text{FBI of previous year}}$$

Conversion From Chain Base to Fixed Base

A CBI is converted to a FBI by the following formula:

$$\text{FBI of a year} = \frac{\text{CBI of the year x FBI of previous year}}{100} \quad \ldots (11)$$

Example 9: From the following data which gives the prices of rice per quintal from 1985 to 190, construct index numbers by (i) fixed base 1985 and (ii) chain base method:

Year	1985	1986	1987	1988	1989	1990
Price	45	50	65	78	84	90

Solution: (i) Fixed base as 1985.

We assign the index number 100 to the year 1985. Now, the table of computations is completed as shown below:

Year	Price	Index Number
1985	45	100
1986	50	$\frac{50}{45}$ x 100 = 111.11
1987	65	$\frac{65}{45}$ x 100 = 144.44
1988	78	$\frac{78}{45}$ x 100 = 173.33
1989	84	$\frac{84}{45}$ x 100 = 186.66
1990	90	$\frac{90}{45}$ x 100 = 200

(ii) Chain Base Method:

Year	Price	Link Relative	Chain Index Number
1985	45	100	100
1986	50	$\frac{50}{45} \times 100 = 111.11$	$\frac{111.11 \times 100}{100} = 111.11$
1987	65	$\frac{65}{50} \times 100 = 130$	$\frac{130 \times 111.11}{100} = 144.44$
1988	78	$\frac{78}{65} \times 100 = 120$	$\frac{120 \times 144.44}{100} = 173.33$
1989	84	$\frac{84}{78} \times 100 = 107.7$	$\frac{107.7 \times 173.33}{100} = 186.66$
1990	90	$\frac{90}{84} \times 100 = 107.14$	$\frac{107.14 \times 186.66}{100} = 200$

Remark: The fixed base index (FBI) is same as the chain base index (CBI) when there is only one item in consideration.

10.6 DEFLATING, SPLICING, SHIFTING OF BASE YEAR:

Shifting Of Base Year

The base year chose in constructing an index number will be become outdated after a long period of time. Also we may require sometimes comparing a given series with another but having a different base year. This, both these reasons leads us to the process of *shifting of the base year*. In this process we obtain the new index numbers by multiplying the previous indices with a common factor of $\frac{100}{i}$, where i is the index of the new base year.

Example 10 : For the following data of price index numbers with base year 1970, shift the base to 1975.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Price Index No.	100	110	115	125	120	130	140	135	150	170

Solution: Every price index is multiplied by the common factor of

$$\frac{100}{\text{Index of new base year}} = \frac{100}{130}$$

The table of shifting the base year to 1975 is as follows:

Year	Old price Index	New Price Index
1970	100	$\frac{100}{130}$ x 100 = 77
1971	110	$\frac{100}{130}$ x 110 = 84
1972	115	$\frac{100}{130}$ x 115 = 88
1973	125	$\frac{100}{130}$ x 125 = 96
1974	120	$\frac{100}{130}$ x 120 = 92
1975	130	100
1976	140	$\frac{100}{130}$ x 140 = 108
1977	135	$\frac{100}{130}$ x 135 = 104
1978	150	$\frac{100}{130}$ x 150 = 115
1979	170	$\frac{100}{130}$ x 170 = 131

Splicing

The procedure of combining two or more overlapping series and revising the index numbers to get a continuous series of index numbers is called as *splicing*. In practice, in a series of data the base year considered may not be relevant after a certain period and is replaced with another year as the base. But the continuity has to be maintained regarding the time series. In such cases, splicing is a very helpful tool.

We find a common factor of $\frac{i}{100}$, where i : stands for the index number of the overlapping period from where the continuity is to be formed. This common factor is then multiplied (or divided depending upon which series is to be spliced) to all the remaining series where the splicing is demanded.

Example 11: For the following data, splice the index *B* to index *A* to obtain up-to-date continuous index numbers:

Year	1972	1973	1974	1975	1976
Index *A*	100	120	130	135	150
Year	1976	1977	1978	1979	1980
Index *B*	100	110	130	135	145

Solution: Using the formula for splicing an index number the table of computation is completed as shown below:

Year	Index *A*	Index *B*	Splicing of *B* to *A*
1972	100		
1973	120		
1974	130		
1975	135		
1976	150	100	$\frac{150 \times 100}{100} = 150$
1977		110	$\frac{150 \times 110}{100} = 165$
1978		130	$\frac{150 \times 130}{100} = 195$
1979		135	$\frac{150 \times 135}{100} = 202.5$
1980		145	$\frac{150 \times 145}{100} = 217.5$

Deflating

As discussed earlier in this chapter, index numbers are very useful in finding the real income of an individual or a group of them, which facilitates the different managements to decide their wage policies. The process of measuring the actual income vis-a-vis the changes in prices is called as *deflation*.

The formula for computing the real income is as follows:

$$\text{Real Income of a year} = \frac{\text{Income for the year}}{\text{Price Index of that year}} \times 100$$

Example 12

Calculate the real income for the following data:

Year	1990	1991	1992	1993	1994	1995
Income in Rs.	800	1050	1200	1600	2500	2800
Price Index	100	105	115	125	130	140

Solution: The real income is calculated by the formula:

$$\text{Real income} = \frac{\text{Income for the year}}{\text{price index of that year}} \times 100$$

The table of computation of real income's is completed as shown below:

Year	Income in Rs.	Price Index	Real Income
1990	800	100	800
1991	1050	105	$\frac{1050}{105} \times 100 = 1000$

1992	1200	115	$\frac{1200}{115} \times 100 = 1043$
1993	1600	125	$\frac{1600}{125} \times 100 = 1280$
1994	2500	130	$\frac{2500}{130} \times 100 = 1923$
1995	2800	140	$\frac{2800}{140} \times 100 = 2000$

10.7 PROBLEMS IN CONSTRUCTING INDEX NUMBERS:

There are various problems in constructing an index number. Some of them are discussed below with their remedies:

(1) Purpose of an index number : Any activity to be performed requires a pre defined purpose, so do the construction of an index number. Depending upon the requirement a suitable index number can be constructed and the necessary procedure is followed. The procedure involves selecting the supportive data, its base year and type of index number to be used.

(2) Selection of a suitable base year : As mentioned earlier, the construction of index number is done with a reference year called as base year. Care should be taken that the base year selected is not the year of any *irregular variations* in the variable. Another important point is that the base year selected, should be contemporary with the available data. In absence of this, the comparison is invalid, as a long period forces appreciable changes in the tastes, customs and habits of people.

(3) Selection of data: For constructing an index number, all the items of a group whose change in prices is to be represented need not be taken. A reasonably large (this is a relative term here) sample of items is enough. The sample should not be too large so that it affect the cost and time constraints and should not be so small that it does not represent the qualities of the group under consideration. A standardized sample covering all varieties serves our purpose.

(4) Obtaining Price quotations: As the sample contains different varieties from different places, their prices also vary. So it becomes difficult to obtain the price quotations. This is overcome by appointing an unbiased and reliable agency to quote the prices.

(5) Selection of proper average: An index number is also a type of average. Generally arithmetic mean and geometric mean are used as to construct index numbers. Geometric mean though ideal than arithmetic mean is not a popular average because of its rigorous calculations.

10.8 DEMERITS OF INDEX NUMBERS

(1) There are numerous types and methods of constructing index numbers. If an appropriate method is not applied it may lead to wrong conclusions.

(2) The sample selection may not be representative of the complete series of items.

(3) The base period selection also is personalized and hence may be biased.

(4) Index number is a quantitative measure and does not take into account the qualitative aspect of the items.

(5) Index numbers are approximations of the changes, they may not accurate.

10.9 LET US SUM UP:

In this chapter we have learn:

10.10 UNIT END EXERCISES

1. Define Index Numbers.
2. Write a short note on the importance of Index Numbers.
3. "Index Numbers are the Economical barometers". Discuss this statement with examples.
4. Discuss the steps to construct Index Numbers.
5. What are the problems in constructing an Index Number?
6. Define the terms: (i) Deflation, (ii) Splicing and (iii) Shifting of Base Year. Write short notes on their significance.
7. Define Cost of Living Index Number and explain its importance.
8. What do you mean by (i) Chain Based Index Number and (ii) Fixed Base Index Number? Distinguish between the two.
9. Define (i) Laspeyre's Index Number, (ii) Paasche's Index Number and (iii) Fisher's Index Number. What is the difference between the three? Which amongst them is called as the ideal Index Number? Why?
10. What are the demerits of Index Numbers?
11. From the following data, construct the price index number by simple aggregative method:

Commodity	Unit	Price in 1990	1991
A	Kg	14	18
B	Kg	6	9
C	Litre	5	8
D	Litre	12	20

12. From the following data, construct the price index number for 1995, by simple aggregative method, with 1994 as the base:

Commodity	Unit	Price in 1994	1995
Rice	Kg	8	10
Wheat	Kg	5	6.5
Oil	Litre	10	13
Eggs	Dozen	4	6

13. From the following data, construct the price index number for 1986, by average of price relatives method:

Commodity	Unit	Price in 1985	1986
Banana	Dozen	4	5
Rice	Kg	5	6
Milk	Litre	3	4.5
Slice Bread	One Packet	3	4

14. From the following data, construct the price index number, by method of average of price relatives:

Commodity	Unit	Price in 1988	1990
A	Kg	6	7.5
B	Kg	4	7
C	Kg	10	14
D	Litre	8	12
E	Litre	12	18

15. From the following data, construct the price index number for 1998, by (i) simple aggregative method and (ii) simple average of price relatives method, with 1995 as the base:

Commodity	Unit	Price in	
		1995	1998
Rice	Kg	12	14
Wheat	Kg	8	10
Jowar	Kg	7	9
Pulses	Kg	10	13

16. From the following data, construct the weighted price index number:

Commodity	A	B	C	D
Price in 1985	10	18	36	8
Price in 1986	12	24	40	10
Weight	40	25	15	20

17. From the following data, construct the index number using (i) simple average of price relatives and (ii) weighted average of price relatives:

Commodity	Weight	Price in	
		1988	1990
Rice	4	8	10
Wheat	2	6	8
Pulses	3	8	11
Oil	5	12	15

18. From the data given below, construct the Laspeyre's index number:

Commodity	1975		1976
	Price	Quantity	Price
A	5	10	8
B	6	15	7.5
C	2	20	3
D	10	14	12

19. From the following data given below, construct the (i) Laspeyre's index number, (ii) Paasche's index number and hence (iii) Fisher's index number.

Commodity	1980		1990	
	Price	Quantity	Price	Quantity
A	6	15	9	21
B	4	18	7.5	25
C	2	32	8	45
D	7	20	11	29

20. From the following data given below, construct the (i) Laspeyre's index number, (ii) Paasche's index number and (iii) Fisher's index number.

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
Cement	140	200	167	254
Steel	60	150	95	200
Coal	74	118	86	110
Limestone	35	50	46	60

21. From the following data given below, construct the Kelly's index number:

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	2	8	4	14
B	6	14	7	20
C	8.5	10	12	15
D	14	8	19	12
E	22	60	38	85

22. From the following data, construct the aggregative price index numbers by taking the average price of the three years as base.

Commodity	Price in 1980	Price in 1981	Price in 1982
A	10	12	16
B	16	19	25
C	5	7	10

23. From the following data, construct the price index number by taking the price in 1978 as the base price:

Commodity	Price in 1978	Price in1979	Price in 1980
A	16	18	24
B	4	6	7.5
C	11	15	19
D	20	28	30

24. For the following data if the Laspeyre's Index number is 133.6, then find the missing quantity:

Commodity	Base Year		Current Year Price
	Price	Quantity	
A	10	12	14
B	16	?	20
C	5	15	8

25. For the following data if the value of I_L = 130.19 and I_P = 142.86, find the missing quantities:

Commodity	1980		1990	
	Price	Quantity	Price	Quantity
Oil	12	15	15	20
Milk	6	10	8	15
Eggs	5	5	8	10

26. From the following data, construct (i) I_L, (ii) I_P, (iii) I_F and (iv) I_K

Commodity	1969		1970	
	Price	Quantity	Price	Quantity
Rice	2	10	3	12
Wheat	1.5	8	1.9	10
Jowar	1	6	1.2	10
Bajra	1.2	5	1.6	8
Pulses	4	14	6	20

27. Construct the cost of living index number for 1980 using the Family Budget Method:

Item	Quantity	Price in	
		1975	1980
A	10	5	7
B	5	8	11
C	7	12	14.5
D	4	6	10
E	1	250	600

28. Construct the cost of living index number for the following data with base year as 1989.

Item	Weight	Price in 1989	1990	1991
Food	4	45	50	60
Clothing	2	30	33	38
Fuel	1	10	12	13
House Rent	3	40	42	45
Miscellaneous	1	5	8	10

29. A survey of families in a city revealed the following information:

Item	Food	Clothing	Fuel	House Rent	Misc.
% Expenditure	30	20	15	20	15
Price in 1987	320	140	100	250	300
Price in 1988	400	150	125	250	320

What is the cost of living index number for 1988 as compared to that of 1987?

30. Construct the consumer price index number for the following industrial data:

Item	Weight	Price Index
Industrial Production	30	180
Exports	15	145
Imports	10	150
Transportation	5	170
Other activity	5	190

31. A person spends Rs. 12,000 a month. If the cost of living index number is 130, find the amount the person spends on Clothing and House rent from the following data:

Item	Expenditure in '000 Rs.	Index No.
Food	3	130
Clothing	?	125
Fuel & Lighting	1	140
House Rent	?	175
Miscellaneous	2	190

32. From the following data which gives the prices of rice per quintal from 1975 to 1980, construct index numbers by (i) fixed base 1975 and (ii) chain base method:

Year	1975	1976	1977	1978	1979	1980
Price	100	105	120	125	130	140

33. From the following data of price of a commodity from 1990 to 1996 given below, construct the index numbers by (i) taking 1990 as base and (ii) chain base method:

Year	1990	1991	1992	1993	1994	1995	1996
Price	42	48	50	54	55	58	64

34. Construct the fixed base index numbers from the following chain base index numbers:

Year	1967	1968	1969	1970	1971	1972	1973
C.B.I.N.	105	112	128	135	150	140	160

35. For the following data of price index numbers with base year 1970, shift the base to 1975.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Price Index No.	100	105	110	120	125	135	130	145	155	170

36. The following data is about the price index numbers with base year 1989. Shift the base to 1995.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Price Index No.	100	110	115	125	130	140	145	135	140	150	155

37. For the following data, splice the index *B* to index to index *A* to obtain up-to-date continuous index numbers:

Year	1972	1973	1974	1975	1976	1977
Index *A*	100	110	125	140	180	225
Year	1977	1978	1979	1980	1981	1982
Index *B*	100	105	115	120	115	125

38. Calculate the real income for the following data:

Year	1988	1989	1990	1991	1992	1993
Income in Rs.	500	550	700	780	900	1150
Price Index	100	110	115	130	140	155

39. Calculate the real income for the following data:

Year	1977	1978	1979	1980	1981	1982
Income in Rs.	250	300	350	500	750	1000
Price Index	100	105	110	120	125	140

40. The per capita income and the corresponding cost of living index numbers are given below. Find the per capita real income:

Year	1962	1963	1964	1965	1966	1967
per capita income	220	240	280	315	335	390
cost of living I.N.	100	110	115	135	150	160

Multiple Choice questions:

1) A composite index number is a number that measures an average relative change ___?

a) in single variable b) in a group of relative variables with respect to a base

c) in prices of commodities d) Both A & B

2) Which of the following is method of constructing Index Numbers ?

a) Aggregative Method b) Relative Method

c) Both A & B d) None of the Above.

3) Index Numbers may be categorized in terms of ___?

a) variables b) constants

c) numbers d) All of the Above

4) Laspeyre's index = 110, Paasche's index = 108, then Dorbis-Bowley index is equal to:

a) 110 b) 108 c) 100 d) 109

4) The ratio of a sum of prices will current period to the sum of prices ill the base period, expressed as a percentage is called:

a) Simple price index number

b) Simple aggregative price index number

c) Weighted aggregative price index number

d) Quantity index number

5) For a particular set of data, Laspeyre's index number is 120 and Paasche's index number is 125, then its Drobish-Bowley's index number is_____.

a) 122 b) 123 c) 123.5 d) 122.5

6) The quantity index number _____ measures changes in level of expenditure.

a) Always b) Sometimes c) Rarely d) Never

7) The Laspeyre's and Paasche's index are examples of:

a) Weighted quantity index only b) Weighted index numbers

c) Aggregate index numbers d) Weighted price index only.

8) The cost of living index number is always

a) Weighted index b) price index

c) quantity index d) None

9) Index number shows _____ changes rather than absolute amount of charge.

a) relative b) percentage

c) both (a) and (b) d) none

10) Index number by family budget is Weighted average of price relative.

a) True b) False

10.7 LIST OF REFERENCES

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K Kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

11

PROBABILITY

Unit Structure

11.0 OBJECTIVES:

After going through this unit, you will able to know:

- The basic terminology of probability.
- How real life the concept of probability used.
- Solve basic example on probability.
- Probability axioms.

11.1 INTRODUCTION:

Some time in daily life certain things come to mind like "I will be success today', I will complete this work in hour, I will be selected for job and so on. There are many possible results for these things but we are happy when we get required result. Probability theory deals with experiments whose outcome is not predictable with certainty. Probability is very useful concept. These days many field in computer science such as machine learning, computational linguistics, cryptography, computer vision, robotics other also like science, engineering, medicine and management.

Probability is mathematical calculation to calculate the chance of occurrence of particular happing, we need some basic concept on random experiment , sample space, and events.

Prerequisites and terminology:

Before we start study of probability, we should know some prerequisites which are required to study probability. Also we have to discuss some basic terminology.

Factorial notation: The product of first n natural number is called factorial of n.

It is denoted by $n!$.

i.e. $n! = n \times (n-1) \times (n-2) \times \ldots\ldots \times 2 \times 1$

For e.g. $4! = 4 \times 3 \times 2 \times 1 = 24$

This can be written as $n! = n \times (n-1)!$

Note: The value of $0! = 1$.

Fundamental principal of counting:

If a certain thing can be done by m ways and other thing are done by *n* ways then the total number of ways in which the two things can be done is $(m \times n)$ ways.

For e.g. Suppose you have 2 pants and 3 shirts. How many choices do you have or how many different ways can you dress?

Then, a tree diagram can be used to show all the choices you can make As you can see in the diagram you have choice of 3 different colors of shirts to wear with brown pant. Similarly, with black pant also you have choice of 3 different colors of shirts to wear.

Thus, you have 2 × 3 = 6 choices.

Permutation: A permutation is an arrangement of all or part of a set of objects, with regard to the order of the arrangement. For example, suppose we have a set of three letters: A, B, and C. we might ask how many ways we can arrange 2 letters from that set. Each possible arrangement would be an example of a **permutation**.

$P_r^n = \frac{n!}{(n-r)!}$; r ≤ n (repetition not allowed)

$= n^r$; r ≤ n (repetition allowed)

For e.g. How many words can be formed using four different alphabets of word "COMBINE?"

Here, Total number of words in "COMBINE" is 7. Word can be formed using four different alphabets i.e. here n = 7 & r = 4

${}^7P_4 = \frac{7!}{(7-4)!} = \frac{7\times6\times5\times4\times3!}{3!} = 7 \times 6 \times 5 \times 4 = 840$

Combination: Combination is the selection of items in which order does not matter.

Formula: $C_r^n = \frac{n!}{r!(n-r)!}$

We can use some properties:

1) $C_n^n = 1,\ C_0^n = 1$

2) $C_1^n = n,\ C_{n-1}^n = n$

3) $C_r^n = C_{n-r}^n$,0 ≤ r ≤ n

For e.g. In how many ways can a committee of 3 men and 2 women be formed out of 10 men and 5 women?

Here the selection of 3 men out of 10 men can be done by ${}^{10}C_3$ and the selection of 2 women can be done by ${}^{5}C_2$.

Total number of ways the committee can be selected = ${}^{10}C_3 \times {}^{5}C_2$

$$= \frac{10\times9\times8}{3\times2\times1} \times \frac{5\times4}{2\times1} =$$

$120 \times 10 = 1200.$

11.2 BASIC CONCEPT OF PROBABILITY:

Random experiment: When experiment can be repeated any number of times under the similar conditions but we get different results on same experiment, also result is not predictable such experiment is called random experiment. For. e.g. A coin is tossed, A die is rolled and so on.

Outcomes: The result which we get from random experiment is called outcomes of random experiment.

Sample space: The set of all possible outcomes of random experiment is called sample space. The set of sample space is denoted by S and number of elements of sample space can be written as$n(S)$. For e.g. A die is rolled, we get $= \{1,2,3,4,5,6\}$, $n(S) = 6$.

Events: Any subset of the sample space is called an event. Or a set of sample point which satisfies the required condition is called an events. Number of elements in event set is denoted by $n(E)$. For example in the experiment of throwing of a dia. The sample space is

S = {1, 2, 3, 4, 5, 6 } each of the following can be event i) A: even number i.e. A = { 2, 4, 6} ii) B: multiple of 3 i.e. B = { 3, 6} iii) C: prime numbers i.e. C = { 2, 3, 5}.

Types of events:

Impossible event: An event which does not occurred in random experiment is called impossible event. It is denoted by $\emptyset$ set. i.e. $n(\emptyset) = 0$. For example getting number 7 when die is rolled. The probability measure assigned to impossible event is Zero.

Equally likely events: when all events get equal chance of occurrences is called equally likely events. For e.g. Events of occurrence of head or tail in tossing a coin are equally likely events.

Certain event: An event which contains all sample space elements is called certain events. i.e. $n(A) = n(S)$.

Mutually exclusive events: Two events A and B of sample space S, it does not have any common elements are called mutually exclusive events. In the experiment of throwing of a die A: number less than 2 , B: multiple of 3. There fore $n(A \cap B) = 0$

Exhaustive events: Two events A and B of sample space S, elements of event A and B occurred together are called exhaustive events. For e.g. In a thrown of fair die occurrence of even number and occurrence of odd number are exhaustive events. There fore $n(A \cup B) = 1$.

Complement event: Let S be sample space and A be any event than complement of A is denoted by $\bar{A}$ is set of elements from sample space S, which does not belong to A. For e.g. if a die is thrown, S = {1, 2, 3, 4, 5, 6} and A: odd numbers, A = {1, 3, 5}, then $\bar{A} = \{2,4,6\}$.

Probability: For any random experiment, sample space S with required chance of happing event E than the probability of event E is define as

$$P(E) = \frac{n(E)}{n(S)}$$

Basic properties of probability:

1) The probability of an event E lies between 0 and 1. i.e. $0 \leq P(E) \leq 1$.
2) The probability of impossible event is zero. i.e. $P(\emptyset) = 0$.
3) The probability of certain event is unity. i.e. $P(E) = 1$.
4) If A and B are exhaustive events than probability of $P(A \cup B) = 1$.
5) If A and B are mutually exclusive events than probability of $P(A \cap B) = 0$.
6) If A be any event of sample space than probability of complement of A is given by $P(A) + P(\bar{A}) = 1 \Rightarrow \therefore P(\bar{A}) = 1 - P(A)$.

Example 1: An unbiased coin is tossed three times. Find the probability that i) no heads turn up, ii) only one head turn up, iii) Atleast one head turn up, iv) At most one head turn up.

Solution: When 3 coins are tossed the sample space is as follows:

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

$$\therefore n(S) = 8$$

i) Event A contained no head turn up.

$A = \{TTT\}, \quad n(A) = 1$

$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{1}{8} = 0.125.$

ii) Event B contained only one head turn up.

$B = \{HTT, THT, TTH\}$

$n(B) = 3$

$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{3}{8} = 0.375.$

iii) Event C contained Atleast one head turn up,

$C = \{HHH, HHT, HTH, THH, HTT, THT, TTH\}$

$\therefore n(C) = 7$

$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{7}{8} = 0.875$

iv) Event D contained At most one head turn up,

$D = \{HTT, THT, TTH, TTT\}$

$\therefore n(D) = 4$

$\therefore P(D) = \frac{n(D)}{n(S)} = \frac{4}{8} = 0.5$

Example 2: If two dice are rolled, find the probability that the sum of the numbers turn up on uppermost faces of the dice is i) even number, ii) a prime number, iii) a perfect square, iv) multiple of 4 v) divisible by 3.

Solution: When 2 dice are rolled the sample space is as follows:

$$S = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,3) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

$$\therefore n(S) = 36$$

i) Event A: sum of the numbers turn up on uppermost faces of the dice is even number.

(i.e. score are 2,4,6,8,10,12)

$n(A) = 18$

$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{18}{36} = 0.5.$

ii) Event B: sum of the numbers turn up on uppermost faces of the dice is a prime number.(i.e score are 2,3,5,7,11)

$n(B) = 15$

$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{15}{36} = 0.4167.$

iii) Event C: sum of the numbers turn up on uppermost faces of the dice is a perfect square.(i.e. score are 4, 9)

$n(C) = 7$

$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{7}{36} = 0.194.$

iv) Event D: sum of the numbers turn up on uppermost faces of the dice is multiple of 4.

(i.e score are 4,8, 12)

$n(D) = 9$

$\therefore P(D) = \frac{n(D)}{n(S)} = \frac{9}{36} = 0.25.$

v) Event E: sum of the numbers turn up on uppermost faces of the dice is divisible by 3.

(i.e score are 3,6,9,12)

$n(E) = 12$

$\therefore P(E) = \frac{n(E)}{n(S)} = \frac{12}{36} = 0.33.$

Example 3: From a well-shuffled pack of cards, a card is drawn at random, find the probability that the card drawn is i) a red card, ii) a king card, iii) a face card, iv)a diamond card.

Solution: When a card is drawn at random, the sample space is

$$n(S) = {}^{52}C_1 = 52$$

i) Event A: drawn a red card.

$n(A) = {}^{26}C_1 = 26$

$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{26}{52} = 0.5$

ii) Event B: drawn a king card,

$n(B) = {}^{4}C_1 = 4$

$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{4}{52} = 0.077$

iii) Event C: drawn a face card,

$n(C) = {}^{12}C_1 = 12$

$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{12}{52} = 0.23$

iv) Event D: drawn a diamond card.

$n(D) = {}^{13}C_1 = 13$

$\therefore P(D) = \frac{n(D)}{n(S)} = \frac{13}{52} = 0.25$

Example 4: A box contains 50 tickets numbered 1 to 50. A ticket is drawn at random from the box. Find the probability that the number on the ticket drawn is i) an odd number, ii) multiple of 5, iii) greater than 35.

Solution: When ticket is drawn at random, the sample space is

$$S = \{1, 2, 3, 4, \ldots\ldots, 50\}$$

$$n(S) = 50$$

i) Event A: To select an odd number.

$$n(A) = 25$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{25}{50} = 0.5$$

ii) Event B: To select a multiple of 5,

$$n(B) = 25$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{10}{50} = 0.2$$

iii) Event C: To select a greater than 35.

$$n(C) = 15$$

$$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{15}{50} = 0.3$$

Example 5: A bag contains 6 white and 4 black balls. Two balls are drawn at random from the bag. Find the probability that i) both are white, ii) both are black, iii) one of each color.

Solution: A bag contains total number of balls = 6 + 4 = 10 balls.

When two balls are drawn at random from the bag, the sample space is

$$n(S) = {}^{10}C_2 = \frac{10 \times 9}{2 \times 1} = 45.$$

i) Event A: to select that both are white balls.

$$n(A) = {}^{6}C_2 = \frac{6 \times 5}{2 \times 1} = 15$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{15}{45} = 0.33$$

ii) Event B: to select that both are black balls.

$$n(B) = {}^{4}C_2 = \frac{4 \times 3}{2 \times 1} = 6$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{6}{45} = 0.13$$

iii) Event C: to select that one ball of each color.

$$n(C) = {}^{6}C_1 \times {}^{4}C_1 = 6 \times 4 = 24$$

$$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{24}{45} = 0.53$$

Example 6: Two cards are drawn from a well-shuffled pack of 52 cards. Find the probability that i) one king and one queen, ii) both of same color,

Solution: When two cards is drawn at random, the sample space is

$$n(S) = {}^{52}C_2 = \frac{52 \times 51}{2 \times 1} = 1326$$

i) Event A: to select that one king card and one queen card.

$$n(A) = {}^{4}C_1 \times {}^{4}C_1 = 16$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{16}{1326} = 0.012$$

ii) Event B: to select that both are of same color cards,

$$n(B) = 2 \times {}^{26}C_2 = 650$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{650}{1326} = 0.49$$

Example 7: A committee of 4 is to be formed from a group of 7 boys and 5 girls. Find the probability that the committee consists of i) all girls, ii) 3 boys and 1 girl, iii) No girls.

Solution: Total numbers = 7 boys + 5 girls = 12

When committee of 4 is to be formed out of 12, the sample space is

$$n(S) = {}^{12}C_4 = \frac{12 \times 11 \times 10 \times 9}{4 \times 3 \times 2 \times 1} = 495$$

i) Event A: to select that the committee contains all girls.

$$n(A) = {}^{5}C_4 = 5$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{5}{495} = 0.01$$

ii) Event B: to select that the committee contains 3 boys and 1 girl

$$n(B) = {}^{7}C_3 \times {}^{5}C_1 = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} \times 5 = 35 \times 5 = 175$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{175}{495} = 0.35$$

iii) Event C: to select that the committee contains no girls.(i.e all are boys)

$$n(C) = {}^{7}C_4 = \frac{7 \times 6 \times 5 \times 4}{4 \times 3 \times 2 \times 1} = 35$$

$$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{35}{495} = 0.07$$

Example 8: A room has three lamp sockets for which bulbs are chosen from a set of 10 bulbs of which 4 are defective. What is the probability that i) room is dark ii) room is lighted?

Solution: To select 3 bulbs from 10 bulbs for sockets, the sample space is

$$n(S) = {}^{10}C_3 = \frac{10\times9\times8}{3\times2\times1} = 120$$

i) Event A: to select that the condition for room is dark.(i.e. no light in room or all are defective bulbs)

$$n(A) = {}^{4}C_3 = \frac{4\times3\times2}{3\times2\times1} = 4$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{4}{120} = 0.033$$

ii) Event B: to select that the condition for room is lighted.(i.e atleast 1 bulb is working in room)

We use here complement probability,

P(room is lighted) = 1 – P (room is dark) = 1- $P(A) = 1 - 0.033 = 0.967$.

Example 9: If letter of the word MISHA can be arranged at random. Find the probability that i) vowel are together, ii) vowels are not together, iii) begins and end with vowels.

Solution: When the letter of the word MISHA is arranged at random, the sample space is

$n(S) = {}^{5}P_5 = 5! = 120$.

i) Event A: to arrange the letter that vowel are together.

$$n(s) = {}^{2}P_2 \times {}^{4}P_4 = 2! \times 4! = 2 \times 24 = 48.$$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{48}{120} = 0.4$$

ii) Event B: to arrange the letter that vowels are not together.

$$n(s) = {}^{4}P_2 \times {}^{3}P_3 = 6\times 3! = 6 \times 6 = 36.$$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{36}{120} = 0.3$$

iii) Event C: to arrange the letter that begins and ends with vowels.

$$n(s) = {}^{2}P_2 \times {}^{3}P_3 = 2! \times 3! = 2 \times 6 = 12.$$

$$\therefore P(C) = \frac{n(C)}{n(S)} = \frac{12}{120} = 0.1$$

Check your Progress:

1. Three coins are tossed, find the probability that i) All are tails, ii) at most 1 tail,

   iii) exactly 2 tails.

2. A card is selected at random from a pack of 52 cards. Find the probability that i) a red card, ii) a heart card, iii) a face card.
3. Two cards are drawn from a well-shuffled pack of 52 cards. Find the probability that i) one black and one red card, ii) Both are hearts card, iii) both are from same suit, iv) both are from different suit.
4. A committee of 3 is to be formed from a group of 6 boys and 4 girls. Find the probability that the committee consists of i) all boys, ii) at least one boy, iii) no boys, iv) exactly two boys.
5. Four men and three women have to stand in a row for a photograph. If they choose their position at random, find the probability that i) Women are together, ii) women are not together.
6. Tickets numbered from 1 to 100 are well-shuffled and a ticket is drawn. What is the probability that the drawn ticket has i) an odd number? ii) number 5 and multiple of 5? iii) a number which is a square?
7. A pair of uniform dice is thrown. Find the probability that the sum of the numbers obtained on uppermost face is i) a two digit number, ii) Divisible by 5, iii) less than 5, iv) A perfect square.
8. A box contains 4 red and 3 yellow and 5 green balls. If two balls are selected at random from the box, what is the probability that i) both are red, ii) one red and one green iii) one yellow ball iv) none red ball.
9. If the letter of the word MANISH be arranged at random, what is the probability that i) vowel are together? ii) Vowel are at extremes? iii) Word begins with M and end with S.
10. A lot of 300 gift articles manufacture in a factory contains 60 defective gift articles. If 3 articles are picked from the lot at random, find the probability that they are non-defective.

11.3 PROBABILITY AXIOMS:

Let S be a sample space. A probability function P from the set of all event in S to the set of real numbers satisfies the following three axioms for all events A and B in S.

i) $P \quad (A) \geq 0$.

ii) $P(\emptyset) = 0$ and $P(S) = 1$.

iii) If A and B are two disjoint sets i.e. $A \cap B = \emptyset$) than the probability of the union of A and B is $P(A \cup B) = P(A) + P(B)$.

Theorem: Prove that for every event A of sample space S, $0 \leq P(A) \leq 1$.

Proof: $S = A \cup \bar{A}$, $\emptyset = A \cap \bar{A}$.

$$\therefore 1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

$$\therefore 1 = P(A) + P(\bar{A})$$

$\Rightarrow P(A) = 1 - P(\bar{A})$ or $P(\bar{A}) = 1 - P(A)$.

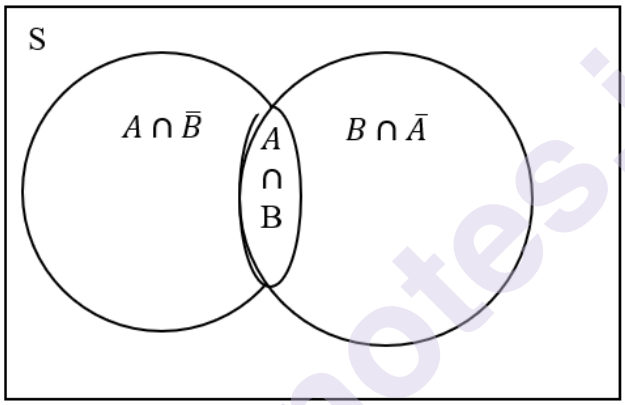
If $P(A) \geq 0$. than $P(\bar{A}) \leq 1$.

$\therefore$ for every event A; $0 \leq P(A) \leq 1$.

11.3.1 Addition theorem of probability:

Theorem: If A and B are two events of sample space S, then probability of union of A and B is given by $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: A and B are two events of sample space S.



Now from diagram probability of union of two events A and B is given by,

$$P(A \cup B) = P(A \cap \bar{B}) + P(A \cap B) + P(B \cap \bar{A})$$

But $P(A \cap \bar{B}) = P(A) - P(A \cap B)$ and $P(B \cap \bar{A}) = P(B) - P(A \cap B)$.

$\therefore P(A \cup B) = P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B)$

$$\therefore P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Note: The above theorem can be extended to three events A, B and C as shown below:

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$

Example 10: A bag contains 4 black and 6 white balls; two balls are selected at random. Find the probability that balls are i) both are different colors. ii) both are of same colors.

Solution: Total number of balls in bag = 4 blacks + 6 white = 10 balls

To select two balls at random, we get

$$n(S) = C(10,2) = 45.$$

i) A be the event to select both are different colors.

$$\therefore n(A) = C(4,1) \times C(6,1) = 4 \times 6 = 24.$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{24}{45} = 0.53.$$

ii) To select both are same colors.

Let Abe the event to select both are black balls

$$n(A) = C(4,2) = 6$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{6}{45}$$

Let B be the event to select both are white balls.

$$n(B) = C(6,2) = 15$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{15}{45}\ .$$

A and B are disjoint event.

∴ The required probability is

$$P(A \cup B) = P(A) + P(B) = \frac{6}{45} + \frac{15}{45} = \frac{21}{45} = 0.467.$$

Example 11: From 40 tickets marked from 1 to 40, one ticket is drawn at random. Find the probability that it is marked with a multiple of 3 or 4.

Solution: From 40 tickets marked with 1 to 40, one ticket is drawn at random

$$n(S) = C(40,1) = 40$$

it is marked with a multiple of 3 or 4, we need to select in two parts.

Let A be the event to select multiple of 3,

i.e. A = { 3,6,9,….,39}

$$n(A) = C(13,1) = 13$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{13}{40}$$

Let B be the event to select multiple of 4.

i.e. B = {4,8,12, …., 40}

$$n(B) = C(10,1) = 10$$

$$P(B) = \frac{n(B)}{n(s)} = \frac{10}{40}.$$

Here A and B are not disjoint.

$A \cap B$ be the event to select multiple of 3 and 4.

i.e. $A \cap B$ = {12,24,36}

$$n(A \cap B) = C(3,1) = 3$$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{3}{40}$$

∴ The required probability is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{40} + \frac{10}{40} - \frac{3}{40} = \frac{20}{40} = 0.5.$$

Example 12: Two cards are drawn at random from a well shuffled pack of cards. Find the probability that i) both are red cards or both are face cards, ii) both are king or queen.

Solution: Two cards are drawn at random from a well shuffled pack of cards than the sample space is $n(S) = {}^{52}C_2 = 1326$.

i) both are red cards or both are face cards

Event A: to select both are red cards

$n(A) = {}^{26}C_2 = 325$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{325}{1326} 0.245$$

Event B: to select both are face cards

$n(B) = {}^{12}C_2 = 66$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{66}{1326} 0.05$$

Event $A \cap B$: To select both red and face cards.

$n(A \cap B) = {}^{6}C_2 = 15$

$$\therefore P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{66}{1326} 0.0045$$

∴ The required probability is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = 0.245 + 0.05 - 0.0045 = 0.2905$$

ii) Both are king or queen.

Event A: to select both are King Cards.

$n(A) = {}^4C_2 = 6$

$$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{6}{1326}$$

Event B: to select both are queen cards.

$n(B) = {}^4C_2 = 6$

$$\therefore P(B) = \frac{n(B)}{n(S)} = \frac{6}{1326}$$

Here event A and Bare mutually exclusive events(i.e. $P(A \cap B) = 0$)

$\therefore$ The required probability is

$$P(A \cup B) = P(A) + P(B) = \frac{6}{1326} + \frac{6}{1326} = \frac{12}{1326} = 0.009$$

Example 13: If the probability is 0.45 that a program development job; 0.8 that a networking job applicant has a graduate degree and 0.35 that applied for both. Find the probability that applied for atleast one of jobs. If number of graduate are 500 then how many are not applied for jobs?

Solution: Let Probability of program development job= $P(A) = 0.45$.

Probability of networking job = $P(B) = 0.8$.

Probability of both jobs = $P(A \cap B) = 0.35$.

Probability of atleast one i.e. to find $P(A \cup B)$.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = 0.45 + 0.8 - 0.35 = 0.9$$

Now there are 500 applications, first to find probability that not applied for job.

$$P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.9 = 0.1$$

Number of graduate not applied for job = $0.1 \times 500 = 50$.

Check your Progress:

1. A card is drawn from pack of 52 cards at random. Find the probability that it is a i) face card or a diamond card, ii)

2. If $P(A) = \frac{3}{8}$ and $(B) = \frac{5}{8}$, $P(A \cup B) = \frac{7}{8}$ than find i) $P(\overline{A \cup B})$ ii) $P(A \cap B)$.

3. In a class of 60 students, 50 passed in computers, 40 passed in mathematics and 35 passed in both. What is the probability that a student selected at random has i) Passed in atleast one subject, ii) failed in both the subjects, iii) passed in only one subject.

4. Two dice are rolled. Find the probability that the sum of the numbers on the uppermost faces is i) an odd number and a prime number, ii) divisible by 2 or 3.

5. A box contains 5 white and 7 black marbles. If 2 marbles are drawn at random. Find the probability that both are of same colors.

11.8 LET US SUM UP:

In this chapter we have learn:

- Basic concept of probability like random experiment, outcomes, sample space, events and types of events.
- Probability Axioms and its basic properties.
- Addition theorem of probability for disjoint events.

11.9 UNIT END EXERCISES:

1. Two coins are tossed. Find the probability that i) Exactly one head turn up ii) atleast one head turn up, iii) no head turn up.

2. An unbiased dice is rolled. Find the probability that i) number less than 3, ii) odd number, iii) a prime number.

3. A card is drawn at random from the pack of well-shuffled 52 cards. Find the probability that it is i) a king, ii) a black, iii) a heart, iv) number between 4 to 9 includes both.

4. A bag contains 5 red and 7 blue marbles and a marble selected at random. Find the probability that a marble is i) red, ii) blue, iii) either red or blue.

5. There are 5 boys and 4 girls and a committee of 4 is to be selected at random. Find the probability that committee contains i) no boys, ii) atleast one boy, iii) exactly 2 boys, iv) atmost 2 boys.

6. Find the probability that a leap year contains 53 Sundays.

7. If two dice are rolled simultaneously, what is the probability of getting the same number on both dice?

8. A card is drawn at random from well shuffled pack of card find the probability that it is red or king card.

9. There are 30 tickets bearing numbers from 1 to 15 in a bag. One ticket is drawn from the bag at random. Find the probability that the ticket bears a number, which is even, or a multiple of 3.

10. In a group of 200 persons, 100 like sweet food items, 120 like salty food items and 50 like both. A person is selected at random find the probability that the person (i). Like sweet food items but not salty food items (ii). Likes neither.

11. A bag contains 7 white balls & 5 red balls. One ball is drawn from bag and it is replaced after noting its color. In the second draw again one ball is drawn and its color is noted. The probability of the event that both the balls drawn are of different colors.

12. A bag contains 8 white & 6 red balls. Find the probability of drawing 2 balls of the same color.

Multiple Choice Questions:

1) Set of all possible outcomes of a random experiments is called

   a) Event b) Experiment

   c) Sample Space d) Space

2) In probability theories, events which can never occur together is known as

   a) mutually exclusive events b) Not exclusive events

   c) mutually exhaustive events d) Complementary event

3) When we throw a dice then what is the probability of getting prime number?

   a) 1/5 b) 1/6 c) ½ d) 1/3

4) A card is drawn at random from a well-shuffled pack of cards. What is the probability that the card drawn is a red?

   a) ½ b) 1/13 c) 2/13 d) 1/4

5) A card is drawn at random from a well-shuffled pack of cards. What is the probability that the card drawn is a diamond?

   a) 1/3 b) 1/13 c) 2/13 d) 1/4

6) Bag contain 10 back and 20 white balls, One ball is drawn at random. What is the probability that ball is white

   a) 1 b) 2/3 c) 1/3 d) 4/3

7) The collection of one or more outcomes from an experiment is called

   a) Probability b) Event

   c) Random Variable d) Random Experiment

8) Which of the following is *not* a correct statement about a probability.

a) It must have a value between 0 and 1

b) It can be reported as a decimal or a fraction

c) A value near 0 means that the event is not likely to occur/happens

d) It is the collection of several experiments.

9) A bag contains 50 tokens numbered 1 to 50. A token is drawn at random. What is probability that number on the token is A multiple of 4?

a) 6/25 b) 6/24 c) 6/23 d) 6/22

10) Any subset of the sample space is called

a) an event b) an element c) superset d) empty set

11.10 LIST OF REFERENCES:

- Schaum's outline of theory and problems on probability and statistics by Murray R. Spiegel.
- Fundamentals of mathematical Statistics by S.C. Gupta and V.K kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

12

PROBABILITY II

Unit Structure

12.0 OBJECTIVES:

After going through this unit, you will able to:

- Conditional probability and its examples.
- Independent events and multiplication theorem of probability.
- Baye's formula of probability.
- Expected value of probability.

12.1 INTRODUCTION:

We have learned basic of probability in previous chapter. Here we are going to discuss bout the conditional probability and its axiom. To find the probability of depended and independed events we used conditional probability. **Conditional probability** is the probability of one event occurring with some relationship to one or more other events. The concept of "randomness" is fundamental to the field of statistics. As mentioned in the probability theory notes, the science of statistics is concerned with assessing the uncertainty of inferences drawn from random samples of data. Now that we've defined some basic concepts related to set operations and probability theory, we can more formally discuss what it means for things to be random. Here we will discuss discrete random variable and its expected value.

12.2 CONDITIONAL PROBABILITY:

In many case we have the occurrence of an event A and are required to find out the probability of occurrence an event B which depend on event A this kind of problem is called conditional probability problems.

Definition: Let A and B be two events. The conditional probability of event B, if an event A has occurred is defined by the relation,

$P(B|A) = \frac{P(B \cap A)}{P(A)}$ if and only if $P(A) > 0$.

In case when $P(A) = 0, P(B|A)$ is not define because $P(B \cap A) = 0$ and $P(B|A) = \frac{0}{0}$ which is an indeterminate quantity.

Similarly, Let A and B be two events. The conditional probability of event A, if an event B has occurred is defined by the relation,

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ If and only if $P(B) > 0$.

Example 1: A pair of fair dice is rolled. What is the probability that the sum of upper most face is 6, given that both of the numbers are odd?

Solution: A pair of fair dice is rolled, therefore $n(S) = 36$.

A to select both are odd number, i.e. A = {(1,1), (1,3), (1,5), (3,1), (3,3), (3,5), (5,1),(5,3), (5,5)}.

$$P(A) = \frac{n(A)}{n(S)} = \frac{9}{36}$$

B is event that the sum is 6, i.e. B = { ((1,5),(2,4), (3,3),(4,2), (5,1)}.

$$P(B) = \frac{n(B)}{n(S)} = \frac{5}{36}$$

$A \cap B$ = { (1,5), (3,3), (5,1)}

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{3}{36}$$

By the definition of conditional probability,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{3/36}{9/36} = \frac{1}{3}.$$

Example 2: If A and B are two events of sample space S, such that $P(A) = 0.85$, $P(B) = 0.7$ and $P(A \cup B) = 0.95$. Find i) $P(A \cap B)$, ii) $P(A|B)$, iii) $P(B|A)$.

Solution: Given that $P(A) = 0.85$, $P(B) = 0.7$ and $P(A \cup B) = 0.95$.

i) By Addition theorem,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$0.95 = 0.85 + 0.7 - P(A \cap B)$

$P(A \cap B) = 1.55 - 0.95 = 0.6.$

ii) By the definition of conditional probability ,

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.6}{0.7} = 0.857.$

iii) $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.6}{0.85} = 0.706$

Example 3: An urn *A* contains 4 Red and 5 Green balls. Another urn *B* contains 5 Red and 6 Green balls. A ball is transferred from the urn *A* to the urn *B*, then a ball is drawn from urn *B*. find the probability that it is Red.

Solution: Here there are two cases of transferring a ball from urn A to B.

Case I: When Red ball is transferred from urn A to B.

There for probability of Red ball from urn A is $P(R_A) = \frac{4}{9}$

After transfer of red ball, urn B contains 6 Red and 6 Green balls.

Now probability of red ball from urn B = $P(R_B|R_A) \times P(R_A) = \frac{6}{12} \times \frac{4}{9} = \frac{24}{108}$.

Case II: When Green ball is transferred from urn A to B.

There for probability of Green ball from urn A is $P(G_A) = \frac{5}{9}$

After transfer of red ball, urn B contains 5 Red and 7 Green balls.

Now probability of red ball from urn B = $P(R_B|G_A) \times P(G_A) = \frac{5}{12} \times \frac{5}{9} = \frac{25}{108}$.

Therefore required probability $= \frac{24}{108} + \frac{25}{108} = \frac{49}{108} = 0.4537.$

Check your progress:

1. A family has two children. What is the probability that both are boys, given at least one is boy?
2. Two dice are rolled. What is the condition probability that the sum of the numbers on the dice exceeds 8, given that the first shows 4?
3. Consider a medical test that screens for a COVID-19 in 10 people in 1000. Suppose that the false positive rate is 4% and the false negative rate is 1%. Then 99% of the time a person who has the condition tests positive for it, and 96% of the time a person who does not have the condition tests negative for it. a) What is the probability that a randomly chosen person who tests positive for the COVID-19 actually has the disease? b) What is the probability that a randomly chosen person who tests negative for the COVID-19 does not indeed have the disease?

4. Out of 200 articles 150 are good and 50 are defective. Find the probability that out of 2 articles are selected at random i) both are good articles. ii) first is good article and second is defective.

5. An unbiased coin is tossed three times and the outcomes of successive tossed are different. Find the probability that the last toss has resulted in tail.

12.3 INDEPENDENT EVENTS:

Independent events: Two events are said to be independent if the occurrence of one of them does not affect and is not affected by the occurrence or non-occurrence of other.

i.e. $P(B/A) = P(B)$ or $P(A/B) = P(A)$.

Multiplication theorem of probability: If A and B are any two events associated with an experiment, then the probability of simultaneous occurrence of events A and B is given by

$$P(A \cap B) = P(A)\, P(B/A)$$

Where $P(B/A)$ denotes the conditional probability of event B given that event A has already occurred.

OR

$$P(A \cap B) = P(B)\, P(A/B)$$

Where $P(A/B)$ denotes the conditional probability of event A given that event B has already occurred.

11.5.1 For Independent events multiplication theorem:

If A and B are independent events then multiplication theorem can be written as,

$$\boldsymbol{P(A \cap B) = P(A)\, P(B)}$$

Proof. Multiplication theorem can be given by,

If A and B are any two events associated with an experiment, then the probability of simultaneous occurrence of events A and B is given by

$$P(A \cap B) = P(A)\, P(B/A)$$

By definition of independent events, $P(B/A) = P(B)$ or $P(A/B) = P(A)$.

$$\therefore \boldsymbol{P(A \cap B) = P(A)\, P(B)}\ .$$

Note:

1) If A and B are independent event then, $\bar{A}$ and $\bar{B}$ are independent event.

2) If A and B are independent event then, $\bar{A}$ and B are independent event.

3) If A and B are independent event then, A and $\bar{B}$ are independent event.

Example 4: From a well-shuffled pack of 52 cards, two cards are drawn at random one after the other without replacement. Find the probability that i) both the cards are same color. ii) the first card is a king and other is a queen.

Solution: i) Here we have to select first black and second also black or first red and second also red. i.e. Event A_1 to select black card and event A_2/A_1 that to select a black card or Event B_1 to select black card and event B_2/B_1 that to select a black card.

By multiplication theorem,

The required probability = $P(A_1) \times P(A_2|A_1) + P(B_1) \times P(B_2|B_1)$

$$= \frac{26}{52} \times \frac{25}{51} + \frac{26}{52} \times \frac{25}{51} = \frac{25}{102} + \frac{25}{102} = \frac{50}{102} = \frac{25}{51} = 0.49$$

ii) Here we have to select first card is a king and then second card is a queen. i.e. Event A to select king card and event B/A that to select a queen card.

By multiplication theorem

$P(A \cap B) = P(A) \times P(B|A)$

$= \frac{4}{52} \times \frac{4}{51} = \frac{4}{663} = 0.006.$

Example 5: Manish and Mandar are trying to make Software for company. Probability that Manish can be success is $\frac{1}{5}$ and Mandar can be success is $\frac{3}{5}$, both are doing independently. Find the probability that i) both are success. ii) Atleast one will get success. iii) None of them will success. iv) Only Mandar will success but Manish will not success.

Solution: Let probability that Manish will success is $P(A) = \frac{1}{5} = 0.2.$

Therefore probability that Manish will not success is $P(\bar{A}) = 1 - P(A) = 1 - 0.2 = 0.8.$

Probability that Mandar will success is $P(B) = \frac{3}{5} = 0.6.$

Therefore probability that Mandar will not success is $P(\bar{B}) = 1 - P(B) = 1 - 0.6 = 0.4.$

i) Both are success i.e. $P(A \cap B)$.

$P(A \cap B) = P(A) \times P(B) = 0.2 \times 0.6 = 0.12$ ∵ A and B are independent events.

ii) Atleast one will get success. i.e. $P(A \cup B)$

By addition theorem,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.2 + 0.6 - 0.12 = 0.68.$$

iii) None of them will success. $P(\overline{A \cup B})$ or $P(\bar{A} \cap \bar{B})$

[By DeMorgan's law both are same]

$P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.68 = 0.32.$

Or

If A and B are independent than $\bar{A}$ and $\bar{B}$ are also independent.

$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = 0.8 \times 0.4 = 0.32.$

iv) Only Mandar will success but Manish will not success. i.e. $P(\bar{A} \cap B)$.

$$P(\bar{A} \cap B) = P(\bar{A}) \times P(B) = 0.8 \times 0.6 = 0.48$$

Example 6: 50 coding done by two students A and B, both are trying independently. Number of correct coding by student A is 35 and student B is 40. Find the probability of only one of them will do correct coding.

Solution: Let probability of student A get correct coding is $P(A) = \frac{35}{50} = 0.7$

Probability of student A get wrong coding is $P(\bar{A}) = 1 - 0.7 = 0.3$

Probability of student B get correct coding is $P(B) = \frac{40}{50} = 0.8$

Probability of student B get wrong coding is $P(\bar{B}) = 1 - 0.8 = 0.2.$

The probability of only one of them will do correct coding.

i.e. A will correct than B will not or B will correct than A will not.

$$P(A \cap \bar{B}) + P(B \cap \bar{A}) = P(A) \times P(\bar{B}) + P(B) \times P(\bar{A}).$$

$$= 0.7 \times 0.2 + 0.8 \times 0.3$$

$$= 0.14 + 0.24 = 0.38$$

Example 7: Given that $P(A) = \frac{3}{7}$, $P(B) = \frac{2}{7}$, if A and B are independent events than find i) $P(A \cap B)$, ii) $P(\bar{B})$, iii) $P(A \cup B)$, iv) $P(\bar{A} \cap \bar{B})$.

Solution: Given that $P(A) = \frac{3}{7}$, $P(B) = \frac{2}{7}$.

i) A and B are independent events,

$$\therefore P(A \cap B) = P(A) \times P(B) = \frac{3}{7} \times \frac{2}{7} = \frac{6}{49} = 0.122$$

ii) $P(\bar{B}) = 1 - P(B) = 1 - \frac{2}{7} = \frac{5}{7} = 0.714.$

iii) By addition theorem,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{7} + \frac{2}{7} - \frac{6}{49} = \frac{29}{49}$$
$$= 0.592.$$

iv) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.592 = 0.408.$

Example 8: The probability that a student A can solve a problem is $\frac{2}{3}$, that B can solve it is $\frac{1}{2}$ and that C solve it is $\frac{3}{4}$. if all of them try it independently, what is the probability that the problem solved?

Solution: Given that $(A) = \frac{2}{3}$, $P(B) = \frac{1}{2}$, $P(C) = \frac{3}{4}$.

∴ The probability of problem will not solve by each of them is,

$P(A') = \frac{1}{3}$, $P(B') = \frac{1}{2}$, $P(C') = \frac{1}{4}$.

We are going to used complementary event that problem is solved is problem is not solved.

i.e. P(Problem is solved) = 1 – P(Problem is not solved)

$P(A \cup B \cup C) = 1 - P(A' \cap B' \cap C')$

$= 1 - P(A')\,P(B')\;P(C')$ ∴(A,B,C are independent their complement are also independent)

$$= 1 - \left(\frac{1}{3} \times \frac{1}{2} \times \frac{1}{4}\right)$$

$$= 1 - \frac{1}{24} = \frac{23}{24}$$

Check your progress:

1. If $P(A) = \frac{2}{5}$, $P(B) = \frac{1}{3}$ and if A and B are independent events, find $(i)\ P(A \cap B),\ (ii)\ P(A \cup B),\ (iii) P(\bar{A} \cap \bar{B})$.

2. The probability that A , B and C can solve the same problem independently are $\frac{1}{3}, \frac{2}{5}\ and\ \frac{3}{4}$ respectively. Find the probability that i) the problem remain unsolved, ii) the problem is solved, iii) only one of them solve the problem.

3. The probability that Ram can shoot a target is $\frac{2}{5}$ and probability of Laxman can shoot at the same target is $\frac{4}{5}$. A and B shot independently. Find the probability that (i) the target is not shot at all, (ii) the target is shot by at least one of them. (iii) the target shot by only one of them. iv) target shot by both.

4. Two cards are drawn one after another from a well-shuffled pack of 52 cards. Find the probability that both the cards are diamonds if the cards are drawn i) with replacement, ii) without replacement.

5. A bag contains 6 white balls and 4 black balls of same shape. One ball is removed at random, its color is noted and is replaced in the box. Then a second ball is drawn. Find the probability that, i) both are black, ii) both are white, iii) first is white and the second one is black.

12.4 BAYES FORMULA:

In 1763, Thomas Bayes put forward a theory of revising the prior probabilities of mutually exclusive and exhaustive events whenever new information is received. These new probabilities are called as posterior probabilities. The generalized formula of bayes theorem is given below:

Suppose $A_1, A_2, \ldots\ldots, A_k$ are k mutually exclusive events defined in B (a collection of events) each being a subset of the sample space S such that $\bigcup_{i=1}^{k} A_i = S$ and $P(A_i) > 0, \forall\, i = 1,2,, \ldots k.$

For Some arbitrary event B, which is associated with A_i such that $P(B) > 0$, we can find out the probabilities $P(B|A_1), P(B|A_2), \ldots\ldots, P(B|A_k)$.

In Baye's approach we want to find the posterior probability of an event A_i given that B has occurred. i.e. $P(A_i|B)$.

By definition of conditional probability, $P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}$

$\because B \in S$ such that $B \cap S = B$.

$$B = B \cap (A_1 \cup A_2 \cup \ldots\ldots \cup A_k)$$

$\bigcup_{i=1}^{k} A_i = S$ and A_i's are disjoint.

i.e. $B = (B \cap E_1) \cup (B \cap E_2) \cup \ldots\ldots \cup (B \cap E_k)$

$$\therefore P(B) = \sum_{i=1}^{k} P(B \cap A_i)$$

$P(B \cap A_i) = P(A_i|B) \times P(B) \Rightarrow P(A_i|B) = \frac{P(B \cap A_i)}{P(B)}$

But $P(B \cap A_i) = P(B|A_i)P(A_i)$ and $P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$.

Therefore we get,

$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{k} P(B|A_i)P(A_i).}$ this known as Baye's formula.

Example 9: There are three bags, first bag contains 2 white, 2 black, 2 red balls; second bag 3 white, 2 black, 1 red balls and third bag 1 white 2 black, 3 red balls. Two balls are drawn from a bag chosen at random. These are found to be one white and I black. Find the probability that the balls so drawn came from the third bag.

Solution: Let B_1 be the first bag, B_2 be the second bag and B_3 be the third bag.

A denotes the two ball are white and black.

First select the bag from any three bags,))

i.e. $P(B_1) = P(B_2) = P(B_3) = \frac{1}{3}$.

Probability of white and black ball from first bag:

$P(A|B_1) = \frac{C(2,1)\times C(2,1)}{C(6,2)} = \frac{4}{15}$.

Probability of white and black ball from second bag:

$$P(A|B_2) = \frac{C(3,1) \times C(2,1)}{C(6,2)} = \frac{6}{15}.$$

Probability of white and black ball from third bag:

$$P(A|B_3) = \frac{C(1,1) \times C(2,1)}{C(6,2)} = \frac{2}{15}$$

By Baye's theorem,

$$P(B_3|A) = \frac{P(B_3)P(A|B_3)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)}$$

$$= \frac{\frac{1}{3}\times\frac{2}{15}}{\frac{1}{3}\times\frac{4}{15}+\frac{1}{3}\times\frac{6}{15}+\frac{1}{3}\times\frac{2}{15}} = \frac{2/45}{12/45} = \frac{1}{6}.$$

Example 10: A company has two factories F_1 and F_2 that produce the same chip, each producing 55% and 45% of the total production. The probability of a defective chip at F_1 and F_2 is 0.07 and 0.03 respectively. Suppose someone shows us a defective chip. What is the probability that this chip comes from factory F_1?

Solution: Let F_i denote the event that the chip is produced by factory. A denote the event that chip is defective.

Given that $P(F_1) = 0.55,\ P(F_2) = 0.45, P(A|F_1) = 0.07, P(A|F_2) = 0.03$.

By Bayes' formula,

$P(F_1|A) = \frac{P(F_1)P(A|F_1)}{P(F_1)P(A|F_1)+P(F_2)P(A|F_2)} = \frac{0.55\times0.07}{0.55\times0.07+0.45\times0.03} = \frac{0.0385}{0.052} = 0.74.$

12.5 EXPECTED VALUE:

In order to understand the behavior of a random variable, we may want to look at its average value. For probability we need to find Average is called

expected value of random variable X. for that first we have to learn some basic concept of random variable.

Random Variable: A probability measurable real valued functions, say X, defined over the sample space of a random experiment with respective probability is called a random variable.

Types of random variables: There are two type of random variable.

Discrete Random Variable: A random variable is said to be discrete random variable if it takes finite or countably infinite number of values. Thus discrete random variable takes only isolated values.

Continuous Random variable: A random variable is continuous if its set of possible values consists of an entire interval on the number line.

Probability Distribution of a random variable: All possible values of the random variable, along with its corresponding probabilities, so that $\sum_{i=1}^{n} P_i = 1$, is called a probability distribution of a random variable.

The probability function always follow the following properties:

i) $P(x_i) \geq 0$ for all value of i.

ii) $\sum_{i=1}^{n} P_i = 1$.

The set of values x_i with their probability P_i constitute a discrete probability distribution of the discrete variable X.

For e.g. Three coins are tossed, the probability distribution of the discrete variable X is getting head.

X= x_i	0	1	2	3
$P(x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Expectation of a random variable (Mean) :

All the probability information of a random variable is contained in probability mass function for random variable, it is often useful to consider various numerical characteristics of that random variable. One such number is the expectation of a random variable.

If random variable X takes values $x_1, x_2, \ldots\ldots x_n$ with corresponding probabilities $P_1, P_2, \ldots\ldots P_n$ respectively, then expectation of random variable X is

$E(X) = \sum_{i=1}^{n} p_i x_i$ where $\sum_{i=1}^{n} P_i = 1$

Example 11: In Vijay sales every day sale of number of laptops with his past experience the probability per day are given below:

No. of laptop	0	1	2	3	4	5
Probability	0.05	0.15	0.25	0.2	0.15	0.2

Find his expected number of laptops can be sale?

Solution: Let X be the random variable that denote number of laptop sale per day.

To calculate expected value, $E(X) = \sum_{i=1}^{n} p_i x_i$

$$E(X) = (0 \times 0.05) + (1 \times 0.15) + (2 \times 0.25) + (3 \times 0.2) + (4 \times 0.15) + (5 \times 0.2)$$

$$E(X) = 2.85 \sim 3$$

Therefore expected number of laptops sale per day is 3.

Example 12: A random variable X has probability mass function as follow:

X= x_i	-1	0	1	2	3
P(x_i)	K	0.2	0.3	2k	2k

Find the value of k, and expected value.

Solution: A random variable X has probability mass function,

$\sum_{i=1}^{n} P_i = 1.$

⇒ k + 0.2 + 0.3 + 2k +2k = 1

⇒5k = 0.5

⇒ k = 0.1

Therefore the probability distribution of random variable X is

X= x_i	-1	0	1	2	3
P(x_i)	0.1	0.2	0.3	0.2	0.2

To calculate expected value, $E(X) = \sum_{i=1}^{n} p_i x_i$

$E(X) = (-1 \times 0.1) + (0 \times 0.2) + (1 \times 0.3) + (2 \times 0.2) + (3 \times 0.2) = 1.2$.

Example 13: A box contains 5 white and 7 black balls. A person draws 3 balls at random. He gets Rs. 50 for every white ball and losses Rs. 10 every black ball. Find the expectation of him.

Solution: Total number of balls in box = 5 white + 7 black = 12 balls.

To select 3 balls at random, $n(s) = C(12,3) = \frac{12 \times 11 \times 10}{3 \times 2 \times 1} = 220.$

Let A be the event getting white ball.

A takes value of 0, 1, 2 and 3 white ball.

Case I : no white ball. i.e. A = 0,

$$P(A=0)=\frac{C(7,3)}{220}=\frac{35}{220}$$

Case II: one white ball i.e. A = 1,

$$P(A=1)=\frac{C(5,1)\times C(7,2)}{220}=\frac{105}{220}$$

Case III: two white balls i.e. A = 2,

$$P(A=2)=\frac{C(5,2)\times C(7,1)}{220}=\frac{70}{220}$$

Case IV: three white balls i.e. A = 3,

$$P(A=3)=\frac{C(5,3)}{220}=\frac{10}{220}$$

Now let X be amount he get from the game.

Therefore the probability distribution of X is as follows:

X= x_i	-30	30	90	150
P(x_i)	$\frac{35}{220}$	$\frac{105}{220}$	$\frac{70}{220}$	$\frac{10}{220}$

To calculate expected value, $E(X)=\sum_{i=1}^{n}p_i x_i$

$E(X)=\left(-30\times\frac{35}{220}\right)+\left(30\times\frac{105}{220}\right)+\left(90\times\frac{70}{220}\right)+\left(150\times\frac{10}{220}\right)=$ Rs. 45.

12.6 LET US SUM UP:

In this chapter we have learn:

- Condition Probability for dependent events.
- Independent events.
- For Independent events multiplication theorem.
- Baye's formula and its application.
- Expected Value for discrete random probability distribution.

12.7 UNIT END EXERCISES:

1. The probability of A winning a race is $\frac{1}{3}$ & that B wins a race is $\frac{3}{5}$. Find the probability that (a). either of the two wins a race. b), no one wins the race.

2. Three machines A, B & C manufacture respectively 0.3, 0.5 & 0.2 of the total production. The percentage of defective items produced by A, B & C is 4 , 3 & 2 percent respectively. for an item chosen at random , what is the probability it is defective.
3. An urn A contains 3 white & 5 black balls. Another urn B contains 5 white & 7 black balls. A ball is transferred from the urn A to the urn B, then a ball is drawn from urn B. find the probability that it is white.
4. A husband & wife appear in an interview for two vacancies in the same post. The probability of husband selection is $\frac{1}{7}$ & that of wife's selection is $\frac{1}{5}$. What is the probability that, a). both of them will be selected. b). only one of them will be selected. c). none of them will be selected?
5. A problem statistics is given to 3 students A,B & C whose chances of solving if are $\frac{1}{2}$, $\frac{3}{4}$ & $\frac{1}{4}$ respectively. What is the probability that the problem will be solved?
6. A bag contains 8 white & 6 red balls. Find the probability of drawing 2 balls of the same color.
7. Find the probability of drawing an ace or a spade or both from a deck of cards?
8. A can hit a target 3 times in a 5 shots, B 2 times in 5 shots & C 3 times in a 4 shots. they fire a volley. What is the probability that a).2 shots hit? b). at least 2 shots hit?
9. A purse contains 2 silver & 4 cooper coins & a second purse contains 4 silver & 4 cooper coins. If a coin is selected at random from one of the two purses, what is the probability that it is a silver coin?
10. The contain of a three urns are : 1 white, 2 red, 3 green balls; 2 white, 1 red, 1 green balls & 4 white, 5 red, 3 green balls. Two balls are drawn from an urn chosen at random. This are found to be 1 white & 1 green. Find the probability that the balls so drawn come from the second urn.
11. Three machines A,B & C produced identical items. Of there respective output 2%, 4% & 5% of items are faulty. On a certain day A has produced 30% of the total output, B has produced 25% & C the remainder. An item selected at random is found to be faulty. What are the chances that it was produced by the machine with the highest output?
12. A person speaks truth 3 times out of 7. When a die is thrown, he says that the result is a 1. What is the probability that it is actually a 1?
13. There are three radio stations A, B and C which can be received in a city of 1000 families. The following information is available on the basis of a survey:

    (a) 1200 families listen to radio station A

    (b) 1100 families listen to radio station B.

(c) 800 families listen to radio station C.

(d) 865 families listen to radio station A & B.

(e) 450 families listen to radio station A & C.

(f) 400 families listen to radio station B & C.

(g) 100 families listen to radio station A, B & C.

The probability that a family selected at random listens at least to one radio station.

14. The probability distribution of a random variable x is as follows.

X	1	3	5	7	9
P(x)	K	2k	3k	3k	K

Find value of (i). K (ii). E(x)

15. A player tossed 3 coins. He wins Rs. 200 if all 3 coins show tail, Rs. 100 if 2 coins show tail, Rs. 50 if one tail appears and loses Rs. 40 if no tail appears. Find his mathematical expectation.

16. The probability distribution of daily demand of cell phones in a mobile gallery is given below.

Find the expected mean .

Demand	5	10	15	20
Probability	0.4	0.22	0.28	0.10

17. If $P(A) = \frac{4}{15}$, $P(B) = \frac{7}{15}$ and if A and B are independent events, find $(i)\ P(A \cap B),\ (ii)\ P(A \cup B),\ (iii) P(\bar{A} \cap \bar{B})$.

18. If $P(A) = \frac{5}{9}$, $P(\bar{B}) = \frac{2}{9}$ and if A and B are independent events, find $(i)\ P(A \cap B),\ (ii)\ P(A \cup B),\ (iii) P(\bar{A} \cap \bar{B})$.

19. If $P(A) = 0.65$, $P(B) = 0.75$ and $P(A \cap B) = 0.45$, where A and B are events of sample space S, find $(i) P(A|B)$, $(ii)\ P(A \cup B)$, $(iii) P(\bar{A} \cap \bar{B})$.

20. A box containing 5 red and 3 black balls, 3 balls are drawn at random from box. Find the expected number of red balls drawn.

21. Two fair dice are rolled. X denotes the sum of the numbers appearing on the uppermost faces of the dice. Find the expected value.

Multiple Choice Questions:

1) _____ variable takes only isolated values.

a) Continuous Random b) Discrete random

c) Possible random d)Mean random

2) Variance of random variable X is

a) $V(X) = E(X^2) + [E(X)]^2$ b) $V(X) = E(X^2) - E(X)$

c) $V(X) = [E(X)]^2 - E(X^2)$ d) $V(X) = E(X^2) - [E(X)]^2$

3) The expected value for the following probability distribution of random variable X is

X	2	4	6
P(X)	0.35	0.4	0.25

a) 3.4 b) 2.8 c) 3.2 d) 3.8

4) If for a random variable X, $E(X) = 1.5$ and $E(X^2) = 9.25$, then V(X) is

a) 6.25 b) 6 c) 7 d) 8

5) If for a random variable X, $V(X) = 3.5$ and $E(X^2) = 19.5$. Then E(X) is

a) 4.5 b) 4 c) 4.79 d) 4.2

6) The probability mass function, has condition

a) $0 \leq P(xi) \leq 1$, for each i b) $0 < P(x_i) \leq 1$, for each i

c) $0 \leq P(x_i) < 1$, for each i d) $0 < P(x_i) = 1$, for each i

7) If event A and B are independent events than $P(A \cap B)$ is

a) $P(A) \times P(B)$ b) $P(A) + P(B)$

c) $P(A') \times P(B)$ d) $P(A) \times P(B')$

8) If event A and B are independent events, P(A) = 0.3 and P(B) = 0.5 than $P(A \cap B)$ is

a) 0.8 b) 0.2 c) 0.15 d) 1.5

9) If P(A) = ¼ , P(B) = 4/5 and event A, B are independent then $P(A \cap B)$ is

a) 1/5 b) 1/20 c) ½ d) 1/16

10) If A and B are any two events associated with an experiment, then the probability of simultaneous occurrence of events A and B is given by

a) P(A) + P(B) b) P(A) – P(B) c) P(A∪B) d) P(A∩B)

12.8 LIST OF REFERENCES:

- Schaum‘s outline of theory and problems on probability and statistics by Murray R. Spiegel.
- Fundamentals of mathematical Statistics by S.C. Gupta and V.K kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

13

DECISION THEORY

Unit Structure

13.0 OBJECTIVES:

After going through this unit, you will able to know:

- The term involving in decision making.
- The different environments of decision maker.
- How to make decision for non-probability with different criteria.

13.1 INTRODUCTION:

Every individual has to make some decisions or other regarding his every day activity like:

i) What we are going have in breakfast?

ii) What we are going to wear today?

iii) Which way we will reach office?

iv) Which movie we are going to watch this weekend?

v) Which mobile modal we have to purchase?

Like these many questions comes in mind and we make decision on it. While taking these decisions we have different options and different conditions. The decisions of routine nature do not involve high risks and are consequently trivial in nature. When business executive make decisions, their decisions affect other people like consumers of the product, shareholders of the business unit, and employees of the organization.

Thus, Decision is a choice whereby a person comes to a conclusion about given circumstances/ situation. It represents a course of behavior or action about what one is expected to do or not to do. Decision- making may, therefore, be defined as a selection of one course of action from two or more alternative courses of action. Thus, it involves a choice-making activity and the choice determines our action or inaction.

Some characteristic problems in decision theory:

1. Production problem: a factory produces several products, and we have to decide how much to manufacture from each product such that e.g. the profit should be maximal, or the aim can be maximal profit with minimal use of energy (or labor) during the production process.

2. Investment problem: to choose a portfolio with maximal yield. Constraint: financial, other points to consider risk factors, duration of the investment etc.

3. Work scheduling: e.g. a supermarket employs a certain number of workers. On each day, depending on the trade a certain number of workers have to work. We have to make a weekly schedule of the available workers such that the total weekly wage of the workers be minimal (wage for Saturday is higher than other days).

4. Buying fighter planes. Points to consider: cost, max. speed, reliability etc

5. Tender evaluation. An international bank wants to replace its computers with new ones. How to decide which offer to accept?

   Points to consider: cost, quality of hardware, service conditions, guarantees, etc. In each case the aim is one action: the best production plan, the highest return, optimal work schedule, finding the best fighter planes for the country, etc.

13.2 COMPONENTS OF DECISION THEORY:

Before learn the decision theory we have to learn some basic components of decision theory so that it is easy to understand the concepts.

Decision Making: Decision-making is the selection based on some criteria from two or more possible alternatives. “George R. Terry”

A decision is an act of choice, wherein an executive forms a conclusion about what must be done in a given situation. A decision represents a course of behaviour chosen from a number of possible alternatives. -—D.E. Mc. Farland

From these definitions, it is clear that decision-making is concerned with selecting a course of action from among alternatives to achieve a predetermined objective.

Decision Maker: The person (Individual or a group) who are responsible to take decision of best alternative is called decision maker. While making decision he can use different criterion or different mathematical models, also he has taken care of different environments and psychology of the persons which involves in this.

Course of Action: Number of alternatives is available for decision making is called course of action or acts or strategies. These are under control known to the decision maker.

State of nature: Events that may follows when a particular decision alternative is selected are called state of nature. The state of nature is mutually exclusive and collectively exhaustive with respect to any decision problem. This is not under the control of decision maker.

Pay-off: In order to compare each combination of action and state of nature we need a payoff (e.g. profit or loss). Typically, this will be a numerical value and it will be clear how we compare the payoffs. For example, we seek to maximize profit or to minimize loss. We shall deal with maximization problems (note that, as one is the negation of the other, there is a duality between maximization and minimization problems). Initially, we shall consider the payoff to be monetary.

Pay-off table: The table consists of all pay-offs tabulated for all courses of action $A_1, A_2, \ldots\ldots, A_n$ under all possible situations i.e. state of nature $S_1, S_2, \ldots\ldots, S_m$. Note that these are mutually exclusive and collectively exhaustive in nature.

State of nature	Course of actions				
	A_1	A_2	------	-----	A_n
S_1	a_{11}	a_{12}	--------	--------	a_{1n}
S_2	a_{21}	a_{22}	------	--------	a_{2n}
S_3	.	.	-------	-------	.
.	.	.	-------	-------	.
.	.	.	-------	--------	.
.	.	.	--------	-------	.
S_m	a_{m1}	a_{m2}	-------	-------	a_{mn}

Number of alternative courses of action is available for making the decision. These are also called actions, acts or strategies. These are under control and known to the decision maker.

2. State of nature: Consequences (or events) that may follow when a particular decision alternative (or strategy) is selected are called states of nature. The states of nature a Number of alternative courses

of action is available for making the decision. These are also called actions, acts or strategies. These are under control and known to the decision maker.

2. State of nature: Consequences (or events) that may follow when a particular decision alternative (or strategy) is selected are called states of nature. The states of nature

13.3 TYPES OF DECISION MAKING CRITERIA:

We can experience several times in decision-making where we don't have the necessary information to decide and keep hesitating. It's often a decision that we have a lot of data with the circumstances and are very specific. There are three types of settings for decision making that we can define. There are also -

- Decision Making in Certain Conditions
- Decision Making in Uncertain Conditions
- Decision Making in Risky Conditions

Decision Making in Certain Conditions

Decision-making under such circumstances ensures that the person who makes a decision has all the complete and appropriate knowledge for the decision to be made. With all the data available, the individual can predict the outcome of the decision. We can easily create a particular decision with confidence by being able to predict the result. Typically, the product that gives the best outcome will be used and carried out.

Decision Making in Uncertain Conditions

When you are unaware of the situation, making a decision is similar to the absence of information to help us decide. The decision-maker doesn't know the future because of inadequate knowledge and can't predict the outcome of any choice he has. The decision-maker will have to judge and decide based on their expertise to decide under certain circumstances. They have to communicate and seek advice from people who have more experience if they do not have those experiences. However, there is a slight risk involved because we cannot predict the outcome, but knowledge from the past will close the gap.

The success or failure of the said company would be determined by the nature of the decisions made in it. So before making an important decision, all the knowledge and alternatives available must be studied. The decision-making process will help a great deal. The atmosphere in which they are made is another aspect that impacts these decisions. In which these choices are made, there are a few different types of environments.

Decision Making Under Risk: The last type of decision-making environment is risky environments. Risk environments are when the

probability of multiple events is tied to a decision. You're never sure about the outcomes of your decision other than calculated guesses. Such decisions are associated with events that could either be very successful or quite disastrous for the organization.

When you're faced with such problems, you will have some data available related to the situation, but it's all a game of probabilities. The past experiences of managers play a huge role, and they often have to take a good look at their past when confronted with such decisions.

The best course of action to take in risky environments is first analyzing the risk of all the alternative actions based on the information available to you.

13.3. DECISION MAKING UNDER UNCERTAINTY (NON-PROBABILITY)

A decision problem, where a decision-maker is aware of various possible states of nature but has insufficient information to assign any probabilities of occurrence to them, is termed as decision-making under uncertainty. A decision under uncertainty is when there are many unknowns and no possibility of knowing what could occur in the future to alter the outcome of a decision.

We feel uncertainty about a situation when we can't predict with complete confidence what the outcomes of our actions will be. We experience uncertainty about a specific question when we can't give a single answer with complete confidence.

Launching a new product, a major change in marketing strategy or opening your first branch could be influenced by such factors as the reaction of competitors, new competitors, technological changes, changes in customer demand, economic shifts, government legislation and a host of conditions beyond your control. These are the type of decisions facing the senior executives of large corporations who must commit huge resources.

The small business manager faces, relatively, the same type of conditions which could cause decisions that result in a disaster from which he or she may not be able to recover. A situation of uncertainty arises when there can be more than one possible consequences of selecting any course of action. In terms of the payoff matrix, if the decision-maker selects A_1, his payoff can be X_{11}, X_{12}, X_{13}, etc., depending upon which state of nature S_1, S_2, S_3, etc., is going to occur.

Different criterion for decision making under uncertainty.

There are a variety of criteria that have been proposed for the selection of an optimal course of action under the environment of uncertainty. Each of these criteria make an assumption about the attitude of the decision-maker.

Maximax Criterion*:* This criterion, also known as the criterion of optimism, is used when the decision-maker is optimistic about future.

Maximax implies the maximisation of maximum payoff. The optimistic decision-maker locates the maximum payoff for each possible course of action. The maximum of these payoffs is identified and the corresponding course of action is selected. This is explained in the following example :

Example : Let there be a situation in which a decision-maker has three possible alternatives A_1, A_2 and A_3, where the outcome of each of them can be affected by the occurrence of any one of the four possible events S_1, S_2, S_3 and S_4. The monetary payoffs of each combination of A_i and S_j are given in the following table:

State of nature	Course of action		
	A_1	A_2	A_3
S_1	1800	1900	1700
S_2	1400	1500	1300
S_3	700	600	500
Maximum	1800	1900	1700
Minimum	700	600	500

The optimal course of action in the above example, Since 1900 is maximum out of the maximum payoffs, based on this criterion, is A_2.

Maximin Criterion: This criterion, also known as the criterion of pessimism, is used when the decision-maker is pessimistic about future. Maximin implies the maximisation of minimum payoff. The pessimistic decision-maker locates the minimum payoff for each possible course of action. The maximum of these minimum payoffs is identified and the corresponding course of action is selected.

In above example, Since 700 is maximum out of the minimum payoffs, the optimal action is A_1.

Regret Criterion: This criterion focuses upon the regret that the decision-maker might have from selecting a particular course of action. Regret is defined as the difference between the best payoff we could have realised, had we known which state of nature was going to occur and the realised payoff. This difference, which measures the magnitude of the loss incurred by not selecting the best alternative, is also known as opportunity loss or the *opportunity cost*.

From the payoff matrix (given in above example), the payoffs corresponding to the actions A_1, A_2, ... An under the state of nature S_j are X_{1i}, X_{2j}, ... X_{nj} respectively. Of these assume that X_{2j} is maximum. Then the regret in selecting A_i, to be denoted by R_{ij} is given by $X_{2j} - X_{ij}$, i = 1 to m. We note that the regret in selecting A_2 is zero. The regrets for various actions under different states of nature can also be computed in a similar way.

The regret criterion is based upon the minimax principle, i.e., the decision-maker tries to minimise the maximum regret. Thus, the decision-maker selects the maximum regret for each of the actions and out of these the action which corresponds to the minimum regret is regarded as optimal. The regret matrix of example can be written as given below:

State of nature	Course of action		
	A_1	A_2	A_3
S_1	100	0	200
S_2	0	100	300
S_3	0	100	200
Maximum	100	100	300

From the maximum regret column, we find that the regret corresponding to the course of action is A_1 and A_2 are minimum. Hence, A_1 and A_2 are optimal.

Hurwicz Criterion: The maximax and the maximin criteria, discussed above, assumes that the decision-maker is either optimistic or pessimistic. A more realistic approach would, however, be to take into account the degree or *index of optimism* or *pessimism* of the decision-maker in the process of decision-making. If a, a constant lying between 0 and 1, denotes the degree of optimism, then the degree of pessimism will be 1 - a. Then a weighted average of the maximum and minimum payoffs of an action, with a and 1 - a as respective weights, is computed. The action with highest average is regarded as optimal.

We note that a nearer to unity indicates that the decision-maker is optimistic while a value nearer to zero indicates that he is pessimistic. If $a = 0.5$, the decision maker is said to be neutralist.

We apply this criterion to the payoff matrix of example 17. Assume that the index of optimism $a = 0.6$.

State of nature	Course of action		
	A_1	A_2	A_3
S_1	1800	1900	1700
S_2	1400	1500	1300
S_3	700	600	500
Maximum	1800	1900	1700
Minimum	700	600	500
$[\text{Max}\times a] + [mini \times (1-a)]$	$1800\times 0.6 + 700 \times 0.4 = 1360$	$1900\times 0.6 + 600 \times 0.4 = 1380$	$1700\times 0.6 + 500 \times 0.4 = 1220$

Since 1380 is maximum, Hence A_2 is maximum, it is optimal.

Laplace Criterion: In the absence of any knowledge about the probabilities of occurrence of various states of nature, one possible way out is to assume that all of them are equally likely to occur. Thus, if there are n states of nature, each can be assigned a probability of occurrence = 1/n. Using these probabilities, we compute the expected payoff for each course of action and the action with maximum expected value is regarded as optimal.

State of nature	Course of action		
	A_1	A_2	A_3
S_1	1800	1900	1700
S_2	1400	1500	1300
S_3	700	600	500
Average	1300	1333.33	1166.67

Since the average for A_2 is maximum, it is optimal.

Example 1: For the following pay-off table, obtained the best decision using i) Maximax criterion, ii) Maximin criterion.

State of nature	Course of action		
	A_1	A_2	A_3
S_1	85	95	70
S_2	40	50	30
S_3	70	60	50

Solution: Select maximum and minimum value from each course of action from pay-off table.

State of nature	Course of action		
	A_1	A_2	A_3
S_1	85	95	70
S_2	40	50	30
S_3	70	60	50
Maximum	85	95	70
Minimum	40	50	30

i) Maximax = Max (Maximum)

= Max (85, 95, 70)

= 95

Which correspond to the course of action A_2.

∴ The best decision is A_2.

ii) Maximin = Max (Minimum)

= Max (40, 50, 30)

= 50

Which correspond to the course of action A_2.

∴ The best decision is A_2.

Example 2: Following payoff tables are given about the demands and different types of product, obtain the best decision using the person of mind set with i) optimistic ii) Pessimistic.

Products	Demands		
	High	Medium	Low
P_1	850	650	300
P_2	1050	700	400
P_3	1200	950	500

Solution: First interchange raw and column of table:

Demands	Products		
	P_1	P_2	P_3
High	850	1050	1200
Medium	650	700	950
Low	300	400	500
Maximum	850	1050	1200
Minimum	300	400	500

i) Optimistic: Maximax = Max(Maximum)

= Max(850, 1050, 1200)

= 1200

Therefore, 1200 belongs to product P_3.

Therefore optimal decision isP_3.

ii) Pessimistic: Minimax = Max(Minimum)

= Max(300, 400, 500)

= 500

Therefore, 300 belongs to product P_3.

Therefore optimal decision isP_3.

Example 3: For the following pay-off table, find the best decision using Minimax regret criterion.

Events	Course of action		
	A_1	A_2	A_3
E_1	120	140	710
E_2	145	150	130
E_3	160	170	150

Solution: First we have to prepare regret table.

Events	Course of action		
	A_1	A_2	A_3
E_1	50	30	0
E_2	5	0	20
E_3	10	0	20
Maximum	50	30	20

For Minimax regret criterion,

Minimax = Min(Maximum)

= Min(50, 30, 20)

=20

Which correspond to the course of action A_3.

∴ The best decision is A_3.

Example 4: For the given pay-off table, state which alternative can be chosen as the best alternative. Using Hurwicz Alpha criterion as $\alpha = 0.7$.

Alternatives	State of nature			
	S_1	S_2	S_3	S_4
A_1	35	25	12	18
A_2	40	20	15	20
A_3	30	30	18	25

Solution: First interchange Raw and column of pay-off table.

State of nature	Alternatives		
	A_1	A_2	A_3
S_1	35	40	30
S_2	25	20	30
S_3	12	15	18
S_4	18	20	25
Maximum	35	40	30
Minimum	12	15	18
[Max$\times a$] + [$mini \times (1 - a)$]	$35 \times 0.7 + 12 \times 0.3 = 28.1$	$40 \times 0.7 + 15 + 0.3 = 32.5$	$30 \times 0.7 + 18 \times 0.3 = 26.4$

Here Maximum value 32.5 which belong to alternative A_2.

Hence the decision is A_2.

Example 5: Form the given pay-off table, obtain the best decision using Laplace criterion.

Participation	Policy		
	P_1	P_2	P_3
High	80	90	100
Medium	60	50	70
Low	50	40	30

Solution: Here we have to find average of each policy.

Participation	Policy		
	P_1	P_2	P_3
High	80	90	100
Medium	60	50	70
Low	50	40	30
Average	190/3 = 63.33	180/3 = 60	200/3 = 66.67

Using Laplace criterion,

Laplace criterion = Max(Average)

= Max(66.33, 60, 66.67)

= 66.67

Here Maximum value 66.67 which belong to alternativeP_3.

Hence the decision is P_3.

Example 6: The following table presents the pay-off returns associated with four alternative types of investment decision.

State of economy	Investment alternative (Rs 10,000)			
	Saving A/C	Fixed Deposit	Mutual Fund	Stock
Recession	450	500	800	-250
Stable	400	550	950	1200
Expansion	500	650	1050	1500

State which can be chosen as the best act using: (a) Maximax, (b) Maximin, (c) Equal likelihood (Laplace), (d) Hurwicz Alpha criterion α=0.4 (e) Minimax regret (savage criterion),.

Solution:

State of economy	Investment alternative (Rs 10,000)			
	Saving A/C	Fixed Deposit	Mutual Fund	Stock
Recession	450	500	800	-250
Stable	400	550	950	1200
Expansion	500	650	1050	1500
Maximum	500	650	1050	1500
Minimum	400	500	800	-250
Average	1350/3=450	1700/3=566.67	2800/933.33	2450/3=816.67
[Max× a] + [$mini \times (1-a)$]	$500 \times 0.4 + 400 \times 0.6 = 440$	$650 \times 0.4 + 500 \times 0.6 = 460$	$1050 \times 0.4 + 800 \times 0.6 = 900$	$1500 \times 0.4 + (-250 \times 0.6) = 450$

(a) Maximax: Max(Maximum) = Max(500, 650, 1050, 1500) = 1500.

Here the value 1500 belong to alternative Stocks.

Hence the decision is invest in Stock.

(b) Maximin = Max(Minimum) = Max(400, 500, 800, -250) = 800.

Here the value 800 belong to alternative Mutual Fund.

Hence the decision is invest in Mutual Fund.

(c) Equal likelihood (Laplace) = Max(Average) = Max(450, 566.67, 933.33, 816.67) = 933.33.

Here the value 933.33 belong to alternative Mutual Fund.

Hence the decision is invest in Mutual Fund.

(d) Hurwicz Alpha criterion α=0.4 = Max(440, 460, 900, 450) = 900.

Here the value 900 belong to alternative Mutual Fund.

Hence the decision is invest in Mutual Fund

(e) Minimax regret (savage criterion),.

Prepared regret table,

State of economy	Investment alternative (Rs 10,000)			
	Saving A/C	Fixed Deposit	Mutual Fund	Stock
Recession	350	300	0	1050
Stable	800	650	250	0
Expansion	1000	850	450	0
Maximum	1000	850	450	1050

Minimax regret (savage criterion) = Min(1000, 850, 450, 1050) = 450.

Here the value 900 belong to alternative Mutual Fund.

Hence the decision is invest in Mutual Fund

13. 4 LET US SUM UP:

In this chapter we have learn

- Basic term requirement of decision theory.
- Different environments of decision making.
- Different methods of decision making under uncertainty.

13.5 UNIT END EXERCISES:

1. Given the following pay-off table, obtain the optimum decision using i) Maximax Criterion, ii)Maximin criterion, iii) Laplace criterion.

State of nature	Course of action		
	A_1	A_2	A_3
S_1	4000	2500	2500
S_2	3500	2500	1200
S_3	2000	3600	2000

2. For the following pay off table. Find the best decision using minimax regret criterion.

State of nature	Course of action		
	A_1	A_2	A_3
S_1	400	250	250
S_2	350	250	120
S_3	200	360	200

3. For the given pay-off table, for different mindset persons obtain the optimal decision using

   i) Optimistic ii) Prismatic.

Evants	Acts		
	A_1	A_2	A_3
E_1	40	20	25
E_2	35	25	12
E_3	20	36	20

4. For the given pay-off table, state which alternative can be chosen as the best alternative. Using Hurwicz Alpha criterion as $\alpha = 0.8$.

Alternatives	State of nature			
	S_1	S_2	S_3	S_4
A_1	350	250	120	180
A_2	400	200	150	200
A_3	300	300	180	250

5. For the given pay-off table, state which alternative can be chosen as the best alternative. Using i) Maximax Criterion, ii)Maximin criterion, iii) Laplace criterion.

Alternatives	State of nature			
	S_1	S_2	S_3	S_4
A_1	650	750	150	180
A_2	400	850	150	200
A_3	300	900	180	250

6. A management is faced with the problem of choosing one of three products for manufacturing.

The potential demand for each product may turn out to be good, moderate or poor. Suggest the management the best product using 1) Maximax, 2) Maximin, 3) Equal likelihood (Laplace), 4) Hurwicz Alpha criterion α = 0.4 5) Minimax regret (savage criterion).

Product	Nature of Demand		
	Good	Moderate	Poor
W	20,000	18,000	10,000
X	15,000	17,000	12,000
Y	12,000	18,000	15,000
Z	21,000	19,000	14,000

7. The research department of consumer products division has recommended the marketing department to launch a soap with 3 different perfumes. The marketing manager has to decide the type of perfume to launch under the following estimated payoff for various levels of sales.

Sales (Units)	Types of perfume		
	I	II	III
20,000	40	60	30
15,000	60	70	20
10,000	30	50	10

Find the best decision using (i)Maximax , (ii) Maximin , (iii) Minimax Regret and (iv) Laplace criteria.

8. Construct a payoff matrix for the following situations and find the best decision using (i) Maximin, (ii) Maximax, (iii) Laplace criteria.

Product	Fixed cost (Rs)	Variable cost (Rs.)
X	500	15
Y	400	12
Z	300	10

The likely demand (units) of products Poor demand 300 Moderate demand 700 High demand 1000 Selling price of each product is Rs. 25.

9. for the following pay-off table, obtained the best product using 1) Maximax, 2) Maximin, 3) Equal likelihood (Laplace), 4) Hurwicz Alpha criterion α = 0.6) Minimax regret (savage criterion).

Evants	Acts		
	A_1	A_2	A_3
E_1	150	420	225
E_2	315	235	125
E_3	220	360	200

10. Suggest the best decision for the given pay-off table using i) Maximax, ii) Maximin, iii) Equal likelihood (Laplace), iv) Hurwicz Alpha criterion α = 0.6 v) Minimax regret (savage criterion).

Evants	Acts		
	A_1	A_2	A_3
E_1	80	90	85
E_2	65	85	92
E_3	70	76	60

13.6 LIST OF REFERENCES:

- Fundamentals of mathematical Statistics by S.C. Gupta and V.K kapoor.
- Basic Statistics by B. L. Agrawal.

❄❄❄❄❄❄❄

14

DECISION THEORY II

Unit Structure

14.0 OBJECTIVES:

After going through this unit, you will able to know:

- The environment of decision making under risk
- How to make decision for probability with different criteria.
- The decision tree technique for multi-stage decision making

14.1 INTRODUCTION:

In the previous chapter we have learn about the non-probability base but with different techniques to make decision. Here we are going to learn if we make decision when probabilities with each state of events are given. To obtain best decision we used pay-off table and given probability.

In our day to day life we take lot of decisions, like purchasing any object or to do investment for that object. In these decisions some are simple in the manner but when there are many possibilities to take the decision at that time risk and uncertainty occurs that which possible condition I should take for the better output. Today by experience we know that few people make decisions after the well deliberated calculations, no matter if the decision situation is in a job situation or in a personal life

In deterministic models, a good decision is judged by the outcome alone. However, in probabilistic models, the decision maker is concerned not only with the outcome value but also with the amount of risk each decision carries. As an example of deterministic versus probabilistic models, consider the past and the future. Nothing we can do can change the past, but everything we do influences and change the future, although the future has an element of uncertainty

14.2 DECISION MAKING UNDER RISK (PROBABILITISTICS):

In case of decision-making under uncertainty the probabilities of occurrence of various states of nature are not known. When these probabilities are known or can be estimated, the choice of an optimal action, based on these probabilities, is termed as decision making under risk.

Risk implies a degree of uncertainty and an inability to fully control the outcomes or consequences of such an action. Risk or the elimination of risk is an effort that managers employ. However, in some instances the elimination of one risk may increase some other risks. Effective handling of a risk requires its assessment and its subsequent impact on the decision process. The decision process allows the decision-maker to evaluate alternative strategies prior to making any decision. The process is as follows:

- The problem is defined and all feasible alternatives are considered. The possible outcomes for each alternative are evaluated.
- Outcomes are discussed based on their monetary payoffs or net gain in reference to assets or time.
- Various uncertainties are quantified in terms of probabilities.
- The quality of the optimal strategy depends upon the quality of the judgments. The decision-maker should identify and examine the sensitivity of the optimal strategy with respect to the crucial factors.

There are two methods to take best decision under risk:

i) Expected Monetary Value (EMV)

ii) Expected Opportunity Loss (EOL)

14.2.1 Expected Monetary value (EMV):

The expected monetary value of a decision is the long run average value of the outcome of that decision. In other words, if we have a decision to make, let's suppose that we could make that exact same circumstances many times. One time a good state of nature may occur and we would have a very positive outcome. Another time we many have a negative outcome because some less favorable state of nature happened. If somehow we could repeat that decision lots and lots of times and determine the outcome for each time and then average all those outcomes then we would have the EMV of the decision alternative.

Step to be followed for EMV calculation:

- First multiple probability with the pay-off table to get EMV table.
- Find the sum of each Course of Action of EMV table.
- Select the maximum EMV from Course of action is the best decision.

For example :

Events	Probability	Course of action		EMV	
		A_1	A_2	A_1	A_2
E_1	0.3	40	30	12	9
E_2	0.5	60	70	30	35
E_3	0.2	80	90	16	18
			Total	58	**62**

Here maximum EMV is 62.

Therefore the best decision is A_2.

Example 1: For the following pay-off table, obtained the best decision using EMV method.

State of nature	Probability	Course of action		
		A_1	A_2	A_3
S_1	0.3	120	140	100
S_2	0.4	140	180	160
S_3	0.3	150	100	150

Solution: Prepare the EMV table.

State of nature	Probability	Course of action			EMV		
		A_1	A_2	A_3	A_1	A_2	A_3
S_1	0.3	120	140	100	48	42	30
S_2	0.4	140	180	160	56	72	64
S_3	0.3	150	100	150	45	30	45
				Total	**149**	144	139

Here maximum EMV is 149.

Therefore the best decision is A_1.

Example 2: A person has to make a choice of purchasing 100 shares of company A or

company B. If the market is high, it will cost him Rs. 30,000, if it is fair, it will cost him Rs. 15,000 and for a low market, it will cost him Rs. 10,000 for company A. While for company B, the corresponding amounts are Rs

35,000, 15,000 and Rs. 12,000 respectively for high, fair and law market condition. The respective probabilities for these market conditions are 0.6, 0.3 and 0.1 respectively. Draw appropriate Decision Tree and advice the person about purchase of shares.

Solution: First covert the data into pay-off and prepare EMV table.

State of Economy	Probability	Company		EMV	
		A	B	A	B
High	0.6	30,000	35,000	18,000	21,000
Fair	0.3	20,000	15,000	6,000	4500
Low	0.1	10,000	12,000	1,000	1200
			Total	25,000	**26,700**

Here maximum EMV is 26,700.

Therefore the best decision is Company B.

EPPI and EVPI:

- Using the information in the decision problem, obtain pay-off table and compute EMV for each decision alternatives (courses of action). Next, compute Expected Pay- -off with Perfect Information (EPPI) and Expected Value of Perfect Information (EVPI).

  EPPI: If the decision maker has perfect information before selecting a course of action, he will select the best alternative (with highest pay-off) corresponding to each state of nature (event). Suppose the dealer can buy market information that can accurately predict market demand (state of nature), he can decide how many units to order. For this purpose, EPPI is computed as follows:

  EPPI = summation of product of probability of each event and maximum pay-off.

  EPPI = ∑ (probability * maximum pay-off for each column)

  Next, EVPI is the difference of expected pay-off without and with perfect information. It is the maximum amount the dealer should spend for obtaining perfect information about the market demand (state of nature).

  EVPI = Expected Value with perfect Information — Expected Value without Perfect Inf

  EVPI = EPPI — Max (EMV).

  EPPI and EVPI:

- Using the information in the decision problem, obtain pay-off table and compute EMV for each decision alternatives (courses of action). Next, compute Expected Pay- -off with Perfect Information (EPPI) and Expected Value of Perfect Information (EVPI).

  EPPI: If the decision maker has perfect information before selecting a course of action, he will select the best alternative (with highest pay-off) corresponding to each state of nature (event). Suppose the dealer can buy market information that can accurately predict market demand (state of nature), he can decide how many units to order. For this purpose, EPPI is computed as follows:

  EPPI = summation of product of probability of each event and maximum pay-off.

  EPPI = ∑ (probability * maximum pay-off for each column)

  Next, EVPI is the difference of expected pay-off without and with perfect information. It is the maximum amount the dealer should spend for obtaining perfect information about the market demand (state of nature).

  EVPI = Expected Value with perfect Information — Expected Value without Perfect Inf

  EVPI = EPPI — Max (EMV).

14.2.2 Expected Pay-off with perfect information (EPPI) and Expected Value of perfect information (EVPI):

EPPI: If the decision maker has perfect information before selecting a course of action, he will select the best alternative (with highest pay-off) corresponding to each state of nature (event). Suppose the dealer can buy market information that can accurately predict market demand(State of nature), he can decide how many units to order. For this purpose, EPPI is computed as follows:

EPPI= Summation of product of probability of each event and maximum pay-off.

EPPI=$\sum(\boldsymbol{Probability \times Maximum\ pay - off\ for\ each\ State\ of\ nature})$

EVPI: EVPI is the difference of expected pay-off without and with perfect information. It is the maximum amount the dealer should spend for obtaining perfect information about the market demand(State of nature)

EVPI = EPPI – Max(EMV).

Example 3: The following is demand distribution of a certain product

No. of unit demanded	10	11	12
Probability	0.3	0.5	0.2

If the product is sold at Rs. 80 per unit with cost price Rs. 60 per unit, obtain the best decision using EMV. Also compute EVPI.

Solution: Here we have to prepare pay-off table,

S.P = Rs. 80

C.P = Rs. 60

Profit per unit = Rs. 20

$$Profit\ function = 20D \qquad D \geq S$$

$$= 20D - 60(S - D) \qquad D < S$$

Demand (D)	Probability	Production(S)			Maximum	EMV		
		10	11	12		10	11	12
10	0.3	200	180	160	200	60	54	48
11	0.5	200	220	200	220	100	110	100
12	0.2	200	220	240	240	40	44	48
	Maximum					200	208	196

Here maximum EMV is 208.

Therefore the best decision is to produce 11 units daily.

EPPI $= 200 \times 0.3 + 220 \times 0.5 + 240 \times 0.2 = 218$

EVPI= EPPI – Max(EMV) = 218 – 208 = 10

Expected Opportunity Loss (EOL):

An alternative approach is to maximize EMV by minimizing expected opportunity loss. First an opportunity loss table is constructed. Then the EOL is computed for each alternative by multiplying the opportunity loss by the probability and adding these together.

EOL is the cost of not picking the best solution.

- First construct an opportunity loss table.
- For each alternative, multiply the opportunity loss by the probability of that loss for each possible outcome and add these together.
- Minimum EOL will always result in the same decision as the Maximum EMV.
- Minimum EOL will always equal EVPI

Example 4: For the given pay-off table, obtained the best optimal decision using EOL method.

Demand	Probability	Alternatives		
		A_1	A_2	A_3
High	0.5	95	80	100
Medium	0.3	45	60	75
Low	0.2	15	20	10

Solution: For the Expected opportunity loss (EOL) first we have to prepare regret table.

Demand	Probability	Regret table			EOL		
		A_1	A_2	A_3	A_1	A_2	A_3
High	0.5	5	20	0	2.5	10	0
Medium	0.3	30	15	0	9	4.5	0
Low	0.2	5	0	10	1	0	2
				Total	12.5	14.5	**2**

Here the minimum EOL is 2.

Therefore the optimal solution is A_3.

14.3 DECISION TREE:

Decision trees are best for projects that involve decisions over time. These results in many possible outcomes. Decision trees are inherently for decision making under risk since we must assign probabilities for each node emanating from a chance node. Decision trees also can incorporate the alternatives into one graphic showing the decisions to be made.

Any problem that can be presented in a decision table can also be graphically illustrated by a decision tree. All decision trees contain decision nodes and state of nature nodes.

- Decision nodes are represented by squares from which one or several alternatives may be chosen.
- State-of-nature nodes are represented by circles out of which one or more state-of-nature will occur.

In drawing the tree, we begin at the left and move to the right. Branches from the squares (decision nodes) represent alternatives, and branches from the circles (state-of-nature node) represent the state of nature.

Decision Tree Analysis:

- Define the problem
- Structure or draw the decision tree

- Assign probabilities to the states of nature
- Estimate payoffs for each possible combination or alternatives and states of nature
- Solve the problem by computing expected monetary values (EMVs) for each state of nature node.

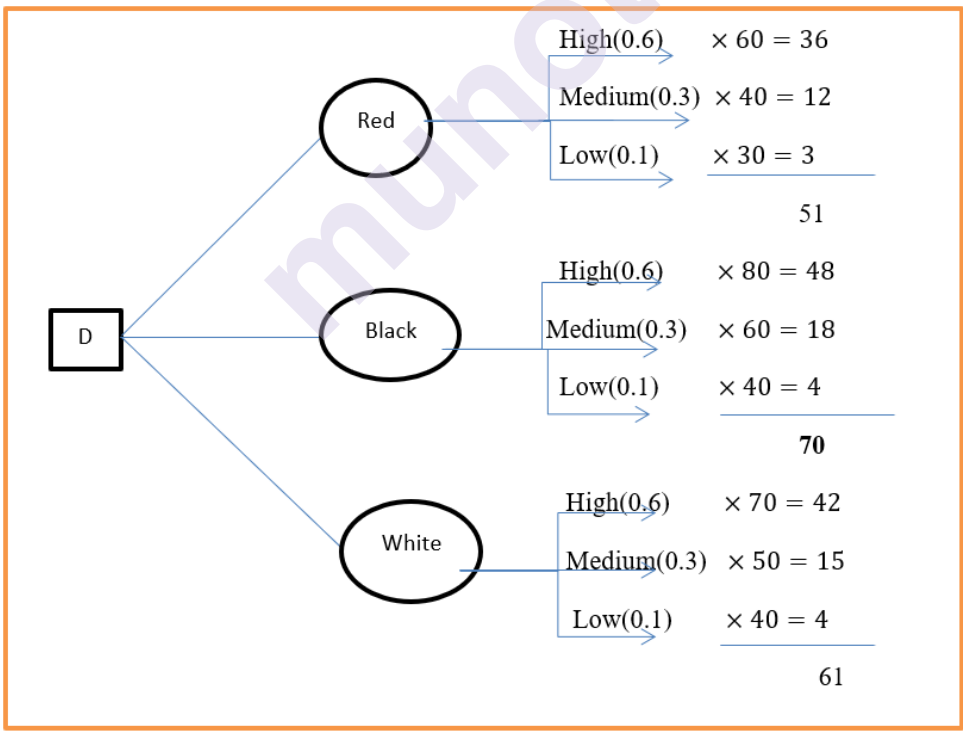
Structure of Decision Trees:

- Trees start from left to right.
- Represent decisions and outcomes in sequential order.
- Squares represent decision nodes
- Circles represent states of nature nodes
- Lines or branches connect the decision nodes and the states of nature

Example 5: Draw the decision tree for the given pay-off table. Also obtained the best decision by EMV method.

Demand	Probability	Colours		
		Red	Black	White
High	0.6	60	80	70
Medium	0.3	40	60	50
Low	0.1	30	40	40

Solution:



Here maximum EMV is 70.

Therefore the best decision is Black Colour.

14.4 LET US SUM UP:

In this chapter we have learn:

- To take decision making under risk.
- Different methods to take decision under risk by EMV and EOL.
- To represent pay-off in decision tree.
- Decision making with decision tree with EMV.

14.5 UNIT END EXERCISES:

1. Write short note on decision tree and procedure of drawing decision tree.
2. Explain Decision making under risk.
3. Explain EMV with one example.
4. Write not on EOL.
5. For the following pay-off table, find optimal decision using EMV method

Course of Action	State of Nature		
	S_1	S_2	S_3
A_1	25	85	95
A_2	40	0	60
A_3	65	30	55

6. The following pay of matrix has been formed by portfolio manager giving pay-offs for different modes of investment under different states of the economy. Decide on the best mode of investment by calculating expected monetary values (EMV).

State of economy	Probability	Investment alternative			
		Gov. F.D	Company F.D	Mutual fund	Shares
Depression	0.25	100	90	50	0
Recovery	0.45	100	110	120	140
Prosperity	0.30	100	120	150	200

7. For the following pay-off table, select the best decision using EOL criteria.

Course of Action	State of Nature		
	S_1	S_2	S_3
A_1	25	85	95
A_2	40	0	60
A_3	65	30	55
Probability	0.5	0.2	0.3

8. Find the best decision by using EOL criterion for the following pair of Matrix.

State of nature	Decisions			
	A_1	A_2	A_3	probability
S_1	20	30	10	0.5
S_2	60	40	30	0.3
S_3	30	70	40	0.2

9. For the following pay of table, suggest the best decision by EOL method.

Course of Action	State of Nature		
	S_1	S_2	S_3
A_1	14	16	10
A_2	12	15	16
A_3	20	18	14
Prob.	0.4	0.3	0.3

10. The following is the demand distribution of a certain product.

No. of units demanded	100	150	200
Probability	0.3	0.45	0.25

If the cost the units is Rs. 250 per unit and selling price is Rs. 340 per unit. Prepare a pay-off table and obtained best decision using EMV.

11. The probability distribution of daily demand of cell phones in a mobile gallery is given below.

Demand	5	10	15	20
Probability	0.4	0.2	0.3	0.1

If the cost the units is Rs. 25000 per unit and selling price is Rs. 34000 per unit. Prepare a pay-off table and obtained best decision using EMV.

12. A company wants to launch a new drink this summer. From the following pay-off table.

    Decide the flavour to be launched using decision tree by EMV method.

Summer condition	Flavour of the soft drink		
	Orange	Mango	Lime
Mild	150	58	59
Moderate	153	151	154
Severe	158	230	198
Very severe	250	268	278

13. Draw a decision tree for the following decision making problem and suggest the best decision.

Course of Action	State of Nature		
	S_1	S_2	S_3
A_1	34	20	18
A_1	14	16	12
Prob.	0.2	0.3	0.5

14. A manufacturer of toys is interested to know whether he should launch a deluxe model or a popular model of a toy. If the deluxe model is launched, the probabilities that the market will be good, fair or poor are given by 0.3,0.4 and 0.3 respectively with payoffs Rs. 1,40,000, Rs. 70,000 and Rs. (-10,000). If the popular model is introduced, the corresponding probabilities are given by 0.4, 0.3 and 0.3 with respective payoffs Rs 1,50,000, Rs. 80,000 and Rs. (-15,000). Decide which model should be launched using decision tree by EMV method.

15. Unique home appliances finds that the cost of holding a cooking ware in stock for a month is Rs. 200. Customer who cannot obtain a cooking ware immediately tends to go to other dealers and he estimates that for every customer who cannot get immediate delivery he loses an average of Rs. 500. The probabilities of a demand of 0, 1, 2, 3, 4, 5 cooking ware in a month are 0.05, 0.1, 0.2, 0.3, 0.2, 0.15 respectively. Determine the optimum stock level of cooking wares. Using EMV criterion.