B.A. II[nd] Year
ECONOMICS

Course Code : ECONA 203
Course Code : 04 (SEC-1)

Statistical Methods – I

Units 1 to 14

By: Prof. Sanju Karol
Dr. Dharam Pal

Centre for Distance and Online Education

Himachal Pradesh University

Summer Hill, Shimla, 171005

Course No. **ECONA 203**

Course title: **Statistical Methods – I**

Nature of Course: **SEC – 1**

Number of credits: **4**

Course Outline

Block	Description
I.	**INTRODUCTION TO STATISTICS** Statistics: Meaning, Scope, Nature, Function, Importance and Limitations of statistics. Types of Data: Primary and Secondary data, Univariate and Bivariate data, qualitative and quantitative data; nominal and ordinal data, Cross-section and Time Series. Sources of Data: Primary and Secondary Data. Diagrammatic and Graphic Presentation of Data.
II.	**CENSUS AND SAMPLE** Collection of Statistical Data: Census and Sample Method, Merits and demerits of census and sampling. Some basic sampling methods: Probability and Non Probability Sampling Methods with merits and demerits. Essentials of sampling, Methods of Selecting Sample, Sampling and Non-Sampling Errors.
III.	**MEASURES OF CENTRAL TENDENCY** Objectives of Averaging, Requisites of a Good Average. Arithmetic Mean, Median, Mode, Geometric Mean and Harmonic Mean. Quartiles, Deciles, Percentiles and Limitations of Averages.
IV.	**DISPERSION** Meaning and significance of dispersion. Measures of dispersion: Range, Quartile Deviation, Mean Deviation, Standard Deviation, Coefficient of Variation, Variance,Absolute and Relative measures of variation - Lorenz Curve.

CONTENTS

Unit-01

STATISTICS: AN INTRODUCTION

STRUCTURE

1.1 INTRODUCTION

Statistics is not a recent field of study; rather, it has existed for as long as human activity itself. However, its scope and applications have expanded significantly over time. In earlier times, statistics was primarily associated with governance and referred to as the "science of statecraft," serving as a tool for administrative functions. Governments of the past maintained records related to population, births, and deaths for governance purposes. The term "statistics" is believed to have originated from the Latin word *status*, the Italian word *statists*, or the German word *Statistik*, all of which relate to the concept of a political state. Today, statistical methods are applied across a

wide range of fields, including agriculture, economics, sociology, and business management. This unit will cover the meaning and definition of statistics, the distinction between descriptive and inferential statistics, the functions and significance of statistics, its limitations, and the skepticism surrounding its use.

1.2 LEARNING OBJECTIVES

By the end of this unit, you will be able to:

- Define the term 'statistics.'
- Differentiate between descriptive and inferential statistics.
- Identify the various functions of statistics.
- Understand the significance of statistical methods across different domains.
- Recognize the limitations of statistical methods.
- explain the reasons for distrust in statistics.

1.3 MEANING OF STATISTICS

The term “statistics” is used in two senses: first in **plural sense** meaning a collection of numerical facts or estimates—the figure themselves. It is in this sense that the public usually think of statistics, e.g., figures relating to population, profits of different units in an industry etc. Secondly, as a **singular sense**, the term ‘statistics’ denotes the various methods adopted for the collection, analysis and interpretation of the facts numerically represented. In singular sense, the term ‘statistics’ is better described as statistical methods. In our study of the subject, we shall be more concerned with the second meaning of the word ‘statistics’. Now let us study in detail about these two approaches.

1.3.1 DEFINITIONS OF STATISTICS

Different writers have defined statistics differently. There are broadly divided into two categories:

1.3.1.1 Definitions in the Plural Sense

1.3.1.2 Definitions in the Singular Sense

1.3.1.1 Definitions in the Plural Sense

In the plural sense, the term Statistics is used for numerical data. Some of its important definitions are:

"statistics are the classified facts representing the conditions of the people in a state.. specially those facts which can be stated in numbers or any tabular or classified arrangement." ---Webster

"numerical statements of facts in any department of enquiry placed in relation to each other.” ---Bowley

“Statistics means quantitative data affected to a marked extent by multiplicity of causes." ---Yule and Kendall

These definitions are quite limited as they restrict the scope of statistics to only those facts or figures that pertain to the state of a population or highlight specific characteristics of data. A broader and more inclusive definition of statistics was provided by Horace Secrist, who described it as:

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

This definition provides a detailed explanation of the essential characteristics that numerical data must possess to be classified as statistics. Let’s examine these characteristics individually.

Characteristics of Statistics in Relation to Numerical Data

i) **Statistics Represent an Aggregate of Facts** – Isolated or unrelated figures are not considered statistics; only aggregated data qualify.

ii) **Statistics Are Influenced by Multiple Factors** – Statistical data are shaped by various interacting elements and forces.

iii) **Statistics Consist of Numerical Facts** – Only numerical data are considered statistics, whereas qualitative descriptions like "small" or "large," "rich" or "poor" do not qualify.

iv) **Statistics Are Measured with a Reasonable Degree of Accuracy** – Data collection should ensure an acceptable level of precision.

v) **Statistics Are Collected Systematically** – Proper planning and appropriate tools must be used by trained individuals to gather data.

vi) **Statistics Are Interrelated** – Data should be comparable and exhibit a meaningful relationship with each other to be classified as statistics.

vii) **Statistics Serve a Specific Purpose** – The objective of data collection should be clearly defined before gathering statistics to ensure relevance and accuracy.

1.3.1.2 Definitions in the Singular Sense

To make informed decisions, numerical data must be collected, organized, presented, analyzed, and interpreted. For this purpose, we rely on specific methods that facilitate these processes. In its singular sense, statistics is defined as a field of study that provides tools for data collection, analysis, and interpretation. However, different scholars have defined statistics in various ways over time.

One notable contributor, Bowley, has provided multiple definitions of statistics, although none of them comprehensively capture its full scope. His definitions reflect the gradual development of the discipline. Some of his notable descriptions include:

i) "Statistics may be called the science of counting."

ii) "Statistics may rightly be called the science of averages."

iii) "Statistics is the science of measurement of social organisms, regarded as a whole in all manifestations."

Croxton and Cowden offer a more straightforward and concise definition: "Statistics may be defined as the collection, presentation, analysis, and interpretation of numerical data."

Seligman provides a more comprehensive yet simple definition: "Statistics is the science that deals with methods for collecting, classifying, presenting, comparing, and interpreting numerical data to gain insights into any area of inquiry."

Fundamental Stages of Statistical Methods

From the various definitions of statistics, it becomes evident that statistical analysis follows a systematic process. This process consists of five key stages:

i) **Data Collection:** Once the research objective is determined, it is crucial to gather relevant information in the form of data. Depending on the study's requirements, data may be obtained from primary sources, secondary sources, or a combination of both.

ii) **Classification and Presentation:** After data collection, researchers must organize the information in a structured manner to facilitate meaningful analysis. Classification involves grouping data based on similarities, making it easier to draw conclusions.

iii) **Tabulation:** The classified data is then presented in a tabular format, making it more comprehensible and suitable for statistical analysis. Tables, along with diagrams and graphs, enhance the visualization of trends and comparisons.

iv) **Data Analysis:** This is the most critical stage of statistical inquiry. The data undergoes statistical processing to extract meaningful patterns and relationships, aiding in decision-making.

v) **Interpretation of Data:** Following analysis, the researcher derives insights regarding the population under study. The quality of interpretation largely depends on the researcher's experience and analytical skills, as it determines the practical applicability of the findings.

From the discussion above, it is evident that the term "statistics" can be used in two contexts. In a plural sense, it refers to numerical data, while in a singular sense, it denotes a set of methods used to analyze data and make well-informed decisions in situations of uncertainty.

Self-Check Exercise 1.1

Q1. Define statistics in the plural sense

Q2. Define statistics in the singular sense

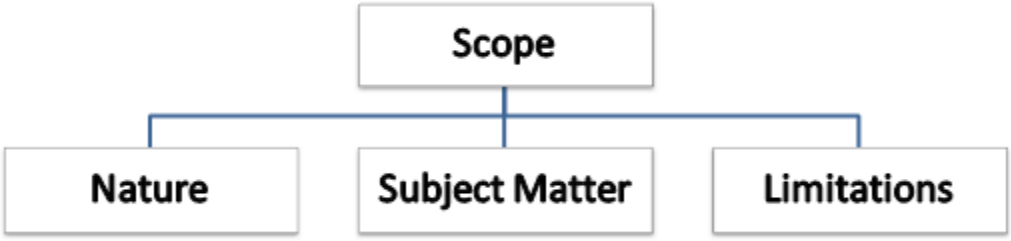
1.4 SCOPE OF STATISTICS

The scope of statistics can be classified in the following parts:

1.4.1 Nature of Statistics

1.4.2 Subject Matter of Statistics

1.4.3 Limitations of Statistics



1.4.1 Nature of Statistics

The study of nature of statistics is to find out whether it is a science or art. As a science, statistics studies numerical data in a systematic manner and as an art, it makes use of the data to solve the problems of real life. Some scholars call it a study of Statistical Methods in preference to Statistics science, because its methods are used in all sciences.

Tippet says, “Statistics is both a science and as art.” It is science as its methods are basically systematic and have general applications. It is an art as its successful application depends to a considerable degree on the skill and special experience of a statistician.

1.4.2 Subject Matter of Statistics

In order to facilitate its study, the subject matter of statistics is divided into two parts namely:

a) Descriptive Statistics

b) Inferential Statistics

(a) Descriptive Statistics

Descriptive statistics refers to the process of analyzing and summarizing data in a meaningful and organized manner. It provides a simple yet effective way to present raw data through numerical calculations, graphs, or tables. This statistical method is essential for making large datasets comprehensible, as it helps in visualizing and understanding patterns in the data. However, it is important to note that descriptive statistics only apply to known data without making predictions or inferences.

Descriptive statistics highlight key characteristics of data by utilizing measures of central tendency such as the mean, median, and mode, along with measures of dispersion like range, variance, and standard deviation. These tools allow for an accurate summary and representation of data using various formats like charts, tables, and graphs. For instance, if we have the exam scores of 1,000 students, descriptive

statistics help in understanding overall performance, distribution, and variability in scores, making the data more interpretable and useful.

(b) Inferential Statistics

Inferential statistics involves analyzing data from a sample to make conclusions or generalizations about a larger population. Its primary objective is to derive insights from a subset of data and apply them to the broader population using probability theory. Common techniques in inferential statistics include hypothesis testing and analysis of variance (ANOVA).

For example, if we aim to study the exam scores of all students in India, collecting data from every student would be impractical. Instead, we can analyze a sample of 1,000 students, ensuring it represents the entire student population. This process, known as sampling, allows researchers to make informed conclusions about the larger group. However, selecting a representative sample is crucial to avoid sampling errors, which could lead to inaccurate results. Inferential statistics primarily rely on two key methods: (1) parameter estimation and (2) statistical hypothesis testing, both of which help in drawing meaningful conclusions from sampled data.

1.4.3 Limitations of Statistics

i) **Neglect of Qualitative Aspects**: Statistical methods focus primarily on quantifiable data and do not account for qualitative characteristics. Aspects such as intelligence, wealth, and health, which cannot be directly measured in numerical terms, are often excluded from statistical analysis. Although attempts are made to convert qualitative information into quantitative data, it remains a challenge. Despite this limitation, statistics is now widely applied across various fields of life and global activities.

ii) **Inability to Analyze Individual Cases**: As stated by Prof. Horace Sacrist, "By statistics, we mean aggregates of facts… and placed in relation to each other." This definition highlights that statistics deals with groups of data rather than individual occurrences. For example, an isolated event like six deaths in an accident or a school's 85% pass rate in a specific year does not qualify as statistical data unless considered as part of a broader dataset. Thus, statistics does not focus on singular instances, no matter how significant they may be.

iii) **Incomplete Representation of a Phenomenon**: Many events occur due to multiple influencing factors, but not all of them can be represented through data. Consequently, statistical conclusions may not always be entirely accurate. For instance, the development of a particular group depends on various social elements, such as parental financial status, education, cultural background, geographic location, and government policies. However, since not all these factors can be quantified, statistical analysis often overlooks crucial qualitative elements, leading to conclusions that are not entirely comprehensive.

iv) **Potential for Misinterpretation**: As W.I. King points out, "One of the shortcomings of statistics is that they do not bear on their face the label of their quality." While data and methodologies can be scrutinized, inaccuracies may arise due to errors by inexperienced researchers or deliberate bias. Since statistics is a delicate

science, it can be easily misused by dishonest individuals. Therefore, data should be handled carefully to avoid misleading conclusions, which could have serious consequences.

v) **Lack of Exact Laws**: Unlike scientific laws, statistical principles—such as the Law of Inertia of Large Numbers and the Law of Statistical Regularity—are based on probability rather than certainty. This means statistical findings are not always as precise as scientific results. For instance, while statistical estimation may predict next year's paddy production, it cannot guarantee a 100% accurate outcome. Instead, statistical laws offer only approximations rather than definitive answers.

vi) **Findings Are True Only on Average**: As discussed earlier, statistical conclusions are often based on estimates using tools like time series analysis, regression, or probability. However, these findings are not universally applicable. For instance, if two sections of students have the same average marks in a statistics exam, it does not imply that all students in both sections scored identically. Individual variations can be significant, and relying solely on averages can sometimes lead to incorrect interpretations.

vii) **Multiple Methods for Analyzing Problems**: Statistics offers various techniques for addressing a single problem, which can sometimes lead to inconsistencies in results. For example, variation can be measured through quartile deviation, mean deviation, or standard deviation, and each method may produce different outcomes. As Croxton and Cowden state, "It must not be assumed that statistics is the only method to use in research, nor should this method be considered the best approach for every problem."

viii) **Statistical Results Are Not Always Conclusive**: According to Prof. L.R. Connor, "Statistics deals only with measurable aspects of things and therefore can seldom give a complete solution to a problem. They provide a basis for judgment but not the whole judgment." Although statistics employs various laws and formulas, its findings are rarely absolute or final. Since it cannot fully resolve all problems, statistical conclusions should be interpreted with caution and applied with sound judgment.

Although numerous laws and formulas are applied in statistics, the results obtained are not always definitive or absolute. Since they do not provide a complete solution to a problem, it is essential to interpret and utilize the findings with careful judgment and discretion.

Self-Check Exercise 1.2

Q1. What is Descriptive Statistics

Q2. What is Inferential Statistics

Q3. List out the limitations of statistics.

1.5 FUNCTIONS OF STATISTICS

The main functions of statistics are as follows:

i) **Presenting Facts in a Clear Form:** Statistics help in representing information accurately using numerical data. Without statistical analysis, ideas may remain vague or ambiguous. Expressing facts numerically makes them more convincing than qualitative descriptions. For example, instead of stating that unemployment is high in India or that the population is rapidly increasing, a precise statistical statement, such as "the population in 2004 was 15 per cent higher than in 1990," provides a clearer picture.

ii) **Simplified Data Presentation:** Statistical techniques condense large amounts of data into meaningful and easy-to-understand figures. By using graphs, charts, averages, or coefficients, complex data can be presented in a more accessible form. For instance, instead of analyzing the prices of multiple goods individually, an overall price index provides a comprehensive view of price trends.

iii) **Facilitating Comparisons:** Once data is simplified, it becomes easier to compare and establish relationships between different groups. Mathematical tools such as averages and coefficients help in drawing meaningful comparisons. Since raw numerical figures alone may not convey much significance, statistical methods provide a structured approach to comparative analysis.

iv) **Formulating and Testing Hypotheses:** Statistics play a crucial role in developing and verifying theories. Using statistical techniques, researchers can analyze the impact of different factors, such as the effect of export taxes on tea consumption in other countries or the effectiveness of credit control measures in reducing inflation.

v) **Predicting Future Trends:** Beyond analyzing current data, statistics also help forecast future developments. For example, if the population growth rate is known, future demand for goods can be estimated. Similarly, businesses can use statistical trends to anticipate market conditions and make informed decisions.

vi) **Assisting in Policy Formulation:** Statistical analysis aids in creating effective policies. Government decisions on food imports, housing, and other sectors depend on estimated future needs, which are derived from statistical data. Accurate forecasting is essential for effective policy-making; otherwise, miscalculations could negatively impact planning and resource allocation.

vii) **Expanding Knowledge and Understanding:** According to Prof. Whipple, "Statistics enables one to enlarge his horizon." Engaging with statistical processes enhances logical reasoning and analytical thinking, allowing individuals to make more informed decisions based on evidence.

viii) **Measuring Uncertainty:** While the future is unpredictable, statistics provide tools to reduce uncertainty by analyzing historical data. Techniques such as regression analysis, interpolation, and time series analysis help in making reliable forecasts and identifying trends, thereby aiding decision-making in uncertain situations.

Self-Check Exercise 1.3

Q1. What are the important function of statistics

1.6 USES OR IMPORTANCE OF STATISTICS

Statistics plays a crucial role in various domains, including assessing per capita income, mortality rates, inflation, and population density. Due to its wide-ranging applications, statistics holds significant importance in fields such as industry, economics, mathematics, astronomy, and healthcare. Below are some key areas where statistics is extensively used:

i) **Industry:** Statistics is vital in the industrial sector as business growth depends on factors like demand and supply, customer retention, and market stability. Statistical methods enable businesses to plan production by considering various market influences that impact sales. Additionally, quality control techniques rely on statistical tools to assess product standards. Hence, statistics is integral to the smooth functioning and success of industries.

ii) **Economics:** Economics involves studying the factors that influence a country's economy. Since these factors continuously fluctuate, economists use statistical tools and models to predict changes accurately. Statistics helps in data collection and analysis, facilitating comparisons of economic variables such as demand and supply, trade balance, inflation, and per capita income. Moreover, economists use statistical representations like tables, histograms, and pie charts to present data in a structured and visually appealing manner.

iii) **Mathematics:** Mathematics and statistics are interdependent disciplines. In mathematics, statistical methods such as measures of central tendency, dispersion, estimation, and hypothesis testing are frequently used. Similarly, statistics incorporates mathematical concepts like integration, differentiation, and algebra for data analysis. Both fields work in tandem to ensure data-driven, evidence-based decision-making. For instance, evaluating the feasibility of a large-scale industrial project requires both mathematical and statistical techniques. Due to this interconnection, statistics is often regarded as a branch of applied mathematics.

iv) **Banking:** Statistics plays a fundamental role in the banking sector. Banks manage deposits and provide loans, and their operations are influenced by factors like interest rates and economic stability. Since these variables are subject to change, banks rely on statistical analysis to make informed predictions. Statistical models help banks assess potential loan defaulters, manage risks, and ensure financial stability. For example, in the event of a market crash, statistics can help banks anticipate economic downturns and minimize risks.

v) **Administration:** Government and administrative sectors depend on statistical methods for data collection, analysis, and policy formulation. Statistics is essential in social welfare projects, budget planning, and economic assessments. For instance, when revising employee pay scales or increasing allowances due to rising living costs, statistical tools determine inflation rates and their impact. Additionally, central and state governments use statistical techniques to estimate revenues and expenditures, ensuring effective financial planning.

vi) **Accounting and Auditing:** Accounting relies on accurate data management, and

statistics is essential for decision-making processes such as asset valuation, liquidity assessment, and depreciation calculations. Since financial conditions fluctuate based on market trends, statistical methods help businesses assess risks and make informed financial decisions. For example, electronic assets depreciate rapidly due to technological advancements, and statistical models assist in determining their declining value.

vii) **Natural Sciences:** In scientific research and development, statistical methods are used for experimental analysis. Laboratory and field experiments generate data, which is analyzed using statistical techniques to draw meaningful conclusions. For instance, when testing the effectiveness of a new drug for heart patients, researchers collect health data from patients before and after treatment. Since results may vary across individuals, hypothesis testing helps determine the drug's overall efficacy with statistical confidence.

viii) **Social Sciences:** In social science research, data is often gathered through surveys, questionnaires, and field studies. Statistics ensures the reliability and validity of research instruments and helps in data analysis. For example, social researchers use statistical methods to process survey responses, analyze trends, and draw conclusions. Additionally, statistics assists in determining appropriate sample sizes and selecting suitable sampling techniques for field studies.

ix) **Astronomy:** Astronomy involves calculating celestial body densities, planetary masses, interstellar distances, and other space-related measurements. Precision is critical in this field, as calculation errors can have significant consequences. Statistical methods help astronomers minimize errors and improve accuracy. For example, the least squares method has been used for centuries to determine the motion of stars.

x) **Demography:** Statistics is the backbone of demography, which studies population structures, gender ratios, health indicators, and age distributions. Since it is impractical to collect data from every individual, statistical methods determine sample sizes and appropriate sampling techniques. For instance, if a researcher aims to analyze the mortality rate of cancer patients in a particular region, statistical methods help identify data sources, select representative samples, and estimate overall mortality rates.

Statistics is an indispensable tool across diverse fields, aiding in decision-making, forecasting, and research. Whether in industry, economics, healthcare, or social sciences, statistical methods provide accurate insights that enhance efficiency and effectiveness in various domains.

Self-Check Exercise 1.4

Q1. Explain the importance of statistics in the field of industry.
Q2. Explain the significance of statistics in social sciences.
Q3. Explain the uses of statistics in the field of administration.

1.7 MISUSE AND DISTRUST OF STATISTICS

Sometimes irresponsible, inexperienced people use statistical tools to fulfill their self-motives irrespective of the nature and trend of the data. Because of such various misuses of statistical tools sometimes called an unscrupulous science. There are various misgivings about Statistics . These are as follows :

"Statistics can prove anything" "Statistics is an unreliable science"

"There are three types of lies , namely, lies, damned lies, and statistics."

"An ounce of truth will produce tons of Statistics "

Therefore care and precautions should be taken care for the interpretation of statistical data. "Statistics should not be used as a blind man uses a lamp-post for support instead of illumination"

Self-Check Exercise 1.5

Q1. How can statistics be misused ?

1.8 SUMMARY

In the present time, people must have some knowledge of statistics. In its singular sense, it means statistical methods that include the collection, classification, analysis, and interpretation of data. In the plural sense, it means quantitative information called data. The subject matter of statistics is very vast. We can categorize statistics into two distinct types: descriptive and inferential statistics. Statistics has applications in almost all branches of knowledge as well as all spheres of life. In spite of its wide applicability, it has certain limitations too. Sometimes inexperienced people misuse statistics to fulfill their own motives.

1.9 GLOSSARY

- **Statistics as plural sense**: means numerical facts or observation collected with a definite purpose.
- **Statistics as singular sense**: means science of statistics or statistical methods. It refers to techniques or methods relating to collection, classification, presentation, analysis and interpretation of quantitative data.
- **Descriptive statistics:** refers to a set of methods used to summarize and describe the main features of a dataset, such as its central tendency, variability, and distribution. These methods provide an overview of the data and help identify patterns and relationships.
- **Inferential statistics:** involves the use of a sample to estimate some characteristic in a large population; and to test a research hypothesis about a given population.

1.10 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 1.1

Ans. Q1. Refer to Section 1.3.1.1

Ans. Q2. Refer to Section 1.3.1.2

Self-Check Exercise 1.2

Ans. Q1. Refer to Section 1.4.2 (a)

Ans. Q2. Refer to Section 1.4.2 (b)

Self-Check Exercise 1.3

Ans. Q1. Refer to Section 1.5

Self-Check Exercise 1.4

Ans. Q1. Refer to Section 1.6

Ans. Q2. Refer to Section 1.6

Ans. Q3. Refer to Section 1.6

Self-Check Exercise 1.5

Ans. Q1. Refer to Section 1.7

1.11 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House.
- Jain, T.R. and Aggarwal, S.C. (2022). Business Statistics. V.K Global Publications Pvt. Ltd. New Delhi.

1.12 TERMINAL QUESTIONS

Q1. Define statistics and discuss its functions and limitations.

Q2. What is statistics? Explain the uses of statistics in business world with suitable examples.

Q3. Define statistics. Also explain its scope and uses.

Unit-02

TYPES OF DATA

STRUCTURE

2.1 INTRODUCTION

This unit deals with the meaning of data. Different types of data, i.e., primary and secondary data, univariate and bivariate data, quantitative and qualitative data, categorical and numerical data and cross-section and time series data will be discussed in this unit.

2.2 LEARNING OBJECTIVES

After studying this unit, you will be able to:

- Define data
- List the types of data
- Explain primary and secondary data
- Elucidate quantitative and qualitative data
- Make a difference between categorical and numerical data
- Distinguish between cross-section and time series data
- Clarify the univariate and bivariate data

2.3 MEANING OF DATA

Data serves as the foundational element from which valuable information is extracted. The term "data" is the plural form of "datum," though it is commonly used in both singular and plural contexts. It refers to raw facts or observations, typically related to physical phenomena or business activities. For instance, the sale of a machine tool or an automobile generates various data points that describe those transactions. Data consists of objective measurements that capture attributes (characteristics) of entities such as individuals, locations, objects, and events. These measurements are often represented using symbols, including numbers, words, and codes, which may consist of a combination of numerical, alphabetical, and other characters.

Data can exist in multiple forms, including numerical values, text, audio, and images. Although data in its raw state lacks structure, it can be organized to produce meaningful information. The terms "data" and "information" are frequently encountered in everyday life and are sometimes used interchangeably. Examples of data include dates, weights, prices, costs, quantities of items sold, employees' names, and product names.

Self-Check Exercise 2.1

Q1. What is meant by Data

2.4 TYPES OF DATA

2.4.1 Primary and Secondary Data

Primary data are always collected from the source. It is collected either by the investigator himself or through his agents. There are different methods of collecting primary data. Each method has its relative merits and demerits. The investigator has to choose a particular method to collect the information. While secondary data are those which have already been collected by someone and have gone through the statistical machines. They are usually refined of the raw materials. When statistical methods are applied on primary their shapes become secondary data.

With the above discussion, we can understand that the difference between primary and secondary data is only in terms of degree. That is that the data which is primary in the hands of one becomes secondary in the hands of another.

Differences between Primary and Secondary Data

Basis for comparison	Primary Data	Secondary Data
Definition	Data gathered firsthand for a specific purpose.	Data that has already been collected by another individual or organization.
Originality	These are original, as they are collected directly by the investigator.	These are not original, as they were collected by someone else for a different purpose.
Nature of data	Exists in raw form before processing.	Already processed and available in a refined format.

Reliability and suitability	More reliable and suitable since they are collected for a specific objective.	Less reliable and may not be entirely suitable, as they were gathered for a different purpose.
Time and money	Collection requires significant time and financial resources.	Requires less time and is more cost-effective.
Precaution and editing	No additional precaution or editing is needed, as the data is collected for a defined objective.	Requires careful verification and possible modifications to align with the current study.

2.4.2 Univariate and Bivariate data

Univariate data involves the analysis of a single variable. In this type, data is collected and analyzed for a single variable at a time. The analysis of the data is done using different measures of central tendency like mean, median, mode, range, standard deviation, and variance. It is a basic analysis technique and is used to summarize the data, identify outliers, and detect patterns or trends.

Bivariate data involves the analysis of two variables. In this type, data is collected and analyzed for two variables at the same time. The analysis of the data is done to examine the relationship between the two variables. The analysis is done using different techniques like scatter plot, correlation, and regression analysis. Bivariate data or analysis is used to explore the cause-and-effect relationship between two variables, identify the strength of the relationship between two variables, and identify any patterns or trends.

Differences between Univariate Data and Bivariate Data

Basis for comparison	Univariate Data	Bivariate Data
Variable	• involving two variables	• involving two variables
Causes	• does not deal with causes or relationships	• deals with causes or relationships
purpose	• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
Statistical Analysis	• central tendency - mean, mode, median • dispersion- range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables
Example	How many of the students in the class are female?	Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

2.4.3 Quantitative and Qualitative Data

Qualitative Data refers to the data that provides insights and understanding about a particular problem. It can be approximated but cannot be computed. Hence, the researcher should possess complete knowledge about the type of characteristic, prior to the collection of data. The nature of data is descriptive and so it is a bit difficult to analyze it. Qualitative Data is concerned with the data that is observable in terms of smell, appearance, taste, feel, texture, gender, nationality and so on.

Quantitative Data, as the name suggests is one which deals with quantity or numbers. It refers to the data which computes the values and counts and can be expressed in numerical terms is called quantitative data. In statistics, most of the analysis are conducted using this data. Quantitative data may be used in computation and statistical test. It is concerned with measurements like height, weight, volume, length, size, humidity, speed, age etc. The tabular and diagrammatic presentation of data is also possible, in the form of charts, graphs, tables, etc. Further, the quantitative data can be classified as discrete or continuous data.

Differences between Qualitative Data and Quantitative Data

Basis for comparison	Qualitative data	Quantitative data
Meaning	Qualitative data is the data in which the classification of objects is based on attributes and properties.	Quantitative Data is the type of data which can be measured and expressed numerically.
Research Methodology	Exploratory	Conclusive
Approach	Subjective	Objective
Analysis	Non-Statistical	Statistical
Collection of data	Unstructured	Structured
Determines	Depth of understanding	Level of occurrence
Asks	Why?	How many or How much?
Sample	Small number of non-representative samples	Large number of representative samples
Outcome	Develops initial understanding.	Recommends final course of action.

2.4.4 Categorical and Numerical Data

Categorical data consists of values that can be classified into distinct groups or categories based on labels or names. This classification is typically performed using a matching process that considers data attributes and their similarities. Each element in a categorical dataset, also known as qualitative data, belongs to only one category, and

these categories are mutually exclusive. Categorical data is further divided into two main types:

- **Nominal Data**: This type categorizes data based on labels or names without any inherent order. It functions similarly to a noun and is sometimes referred to as naming data.
- **Ordinal Data**: This category includes data that can be ranked or ordered based on a specific scale. Unlike nominal data, ordinal data follows a meaningful sequence, but the intervals between values may not be uniform.

Numerical Data

Numerical data consists of values represented in numbers rather than descriptive text. This type of data, also known as quantitative data, is collected in numerical form and is used for measuring attributes such as height, weight, and IQ. Numerical data is classified into two types:

- **Discrete Data**: This includes countable values that correspond to natural numbers. Examples include age, the number of students in a classroom, and the number of candidates in an election.
- **Continuous Data**: This type consists of values that fall within a continuous range and cannot be counted individually. These values are typically represented as intervals on a number line. Examples include a student's CGPA and a person's height.

Difference between Categorical Data and Numerical Data

Basis for comparison	Categorical Data	Numerical Data
Definition	Comprises labels or names used for identification.	Consists of numbers rather than words or descriptions.
Alias	Also known as qualitative data, as it is based on attributes and characteristics.	Also referred to as quantitative data, as it represents numerical values used for arithmetic operations.
Examples	**Gender:** • Male • Female • Other	**Test Scores (out of 20):** • Below 5 • 5-10 • 10-15
Types	Nominal data and Ordinal data	Discrete data and Continuous data
Data collection method	• Nominal: Open-ended questions • Ordinal: Multiple-choice questions	Primarily multiple-choice, occasionally open-ended questions.
Data collection tools	Questionnaires, surveys, and interviews	Questionnaires, surveys, interviews, focus groups, and observations

Uses	Used in surveys that require personal details, opinions, and experiences. Common in business research.	Utilized in statistical analysis and arithmetic computations.
Compatibility	Limited compatibility with statistical techniques, making it less preferred for numerical analysis.	Highly compatible with various statistical calculations and methodologies.

2.4.5 Cross-Section Data and Time Series Data

In cross sectional data, there are several variables at the same point in time. The data may be single observations from a sample survey or from all units in a population. Examples of Norwegian cross-section data are the Household Budget Survey for the year 1999. The Manufacturing Statistics for the year 2000, the Population Census for the year 2001.

Data Series Data focuses on observations of the same subject across time, typically at regular intervals. It is information about a single variable collected over an interval of time, such as months, quarters, years, etc. Usually, time series data is useful in business applications. The time interval should be uniform in a time series dataset. Examples of time-series data are National Accounts data (production, private and public consumption, investment, export, import etc.), the Index of Manufacturing Production, the Consumer Price Index and Financial statistics (money stock, exchange rates, interest rates, bank deposits, etc.)

Most often cross-section data are data for micro unit individuals, households, companies, etc. Most often time-series data are macro data or macro-type data, for example time-series for macro-economic variables from the National Accounts.

Difference between Categorical Data and Numerical Data

Basis for comparison	Cross-Section Data	Time Series Data
Definition	Observation on several variables at the same point in time.	information about a single variable collected over a uniform interval of time
Variable	Focus on a several variables.	Focus on a single variable.
Example	Opening price of a number of shares.	Price of a share in every minute.

Self-Check Exercise 2.2

Q1. Write the main differences between Primary Data and Secondary Data.

Q2. What are Quantitative Data and Qualitative Data.

Q3. Distinguish between Cross-Sectional Data and Time Series Data

2.5 SUMMARY

In this unit, we have studied that data is the raw material from which useful information is derived. Different types of data are primary and secondary data, univariate and bivariate data, quantitative and qualitative data, categorical and numerical data and cross-section and time series data have been discussed in this unit.

2.5 GLOSSARY

- **Data:** Data is the raw material from which useful information is derived.
- **Primary data** are always collected from the source. It is collected either by the investigator himself or through his agents.
- **Secondary data** are those which have already been collected by someone and have gone through the statistical machines. They are usually refined of the raw materials.
- **Univariate data:** When one item of information is collected, for example, from each member of a group of people, the data collected is called univariate data.
- **Bivariate data:** Data that contains two items of information such as height and weight of a person is generally called paired data or bivariate data.
- **Quantitative methods** are those which focus on numbers and frequencies rather than on meaning and experience.
- **Qualitative methods** are ways of collecting data which are concerned with describing meaning, rather than with drawing statistical inferences.
- **Categorical data:** Values or observations that can be sorted into groups or categories. Bar charts and pie graphs are used to graph categorical data.
- **Nominal data:** Values or observations can be assigned a code in the form of a number where the numbers are simply labels.
- **Ordinal data:** Values or observations can be ranked (put in order) or have a rating scale attached.
- **Numerical data:** Values or observations that can be measured. And these numbers can be placed in ascending or descending order.
- **Cross-section data**: These are data from units observed at the same time or in the same time period. The data may be single observations from a sample survey or from all units in a population.
- **Time-Series data**: These are data from a unit (or a group of units) observed in several successive periods.

2.7 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 2.1

Ans. Q1. Refer to Section 2.3

Self-Check Exercise 2.2

Ans. Q1. Refer to Section 2.4.1

Ans. Q2. Refer to Section 2.4.3

Ans. Q3. Refer to Section 2.4.5

2.8 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House.
- Jain, T.R. and Aggarwal, S.C. (2022). Business Statistics. V.K Global Publications Pvt. Ltd. New Delhi.

2.9 TERMINAL QUESTIONS

Q1. What are the different types of data? Explain the Qualitative and Quantitative types of data?

Q2. What are the differences between the primary data and secondary data?

Unit-03

SOURCES OF DATA

STRUCTURE

3.1 INTRODUCTION

Dear students, we all understand that data collection plays a crucial role in the research process. The reliability and accuracy of research findings largely depend on obtaining relevant, precise, and sufficient data. Data is broadly categorized into primary and secondary types. In this unit, we will explore various data sources, delve into different methods of gathering primary data, and examine the sources of secondary data in detail. Additionally, we will discuss the advantages and limitations associated with each data source.

3.2 LEARNING OBJECTIVES

After studying this unit, you will be able to:

- define primary and secondary data,
- know about the various methods of primary data collection,
- explore the different sources of secondary data, and
- apply the appropriate method for data collection.

3.3 PRIMARY DATA

Primary data refers to firsthand information that is gathered for the very first time with a specific objective in mind. This type of data is directly collected by the researcher to address a particular issue or research problem under investigation. Unlike secondary data, which is derived from existing sources, primary data is unique, original, and obtained specifically for the study at hand. It is generated through various data collection methods such as surveys, experiments, interviews, and observations.

In experimental research, primary data consists of information collected during the course of an experiment, ensuring that the findings are based on real-time observations and interactions. Additionally, this data can be obtained through direct communication with individuals relevant to the study topic, using structured or unstructured interviews, focus groups, or descriptive research techniques. Because it is collected firsthand, primary data is highly reliable and tailored to meet the specific requirements of the research.

"Data which are gathered originally for a certain purpose are known as primary data." — Horace Secrist

3.3.1 Advantages of Primary Data

Primary data offers several benefits, making it a valuable resource for research purposes:

- One of the major advantages of primary data is that it is original and directly related to the research topic. Since it is collected firsthand, it ensures a high level of accuracy and reliability.
- Another significant benefit is the flexibility in data collection methods. Primary data can be gathered through various approaches such as interviews, telephone surveys, and focus groups. Additionally, advancements in technology enable data collection across different regions and even international borders through emails and postal surveys, allowing for a broader population and extensive geographical coverage.
- Since primary data is collected in real time, it reflects the current situation, offering a more realistic and updated perspective on the subject under investigation. This makes it highly relevant for decision-making and policy formulation.

- The reliability of primary data is generally high because it is obtained directly from the source, ensuring authenticity and reducing the chances of misinterpretation or manipulation.

3.3.2 Disadvantages of Primary Data

Despite its advantages, primary data collection has certain drawbacks, which include:

- One of the primary limitations of collecting primary data is the restricted coverage in cases where data is gathered through direct interviews. To achieve a wider scope, a larger number of researchers or surveyors are required, which may not always be feasible.
- The process of collecting primary data is time-consuming and requires significant effort. By the time the data is collected, analyzed, and presented in the form of a report, the research problem may have evolved, become more complex, or even lost its relevance. This delay can impact the overall effectiveness of the study.
- Designing an appropriate survey or questionnaire is another challenge. The questions must be clear, simple, and easy to comprehend to ensure that respondents provide meaningful and relevant answers. Poorly designed surveys can lead to inaccurate or incomplete data.
- Respondent behavior is another issue in primary data collection. Some participants may delay their responses, while others may provide false or socially desirable answers instead of sharing the actual facts. This can distort the findings and reduce the authenticity of the research.
- The cost associated with primary data collection is often high. It requires extensive resources, including manpower, time, and financial investment. As the number of respondents increases, the cost escalates further, making it an expensive process.
- Incomplete or poorly filled-out questionnaires can have a negative impact on the research outcomes, leading to gaps in data and reducing the overall reliability of the findings.

While primary data is highly beneficial for research due to its accuracy and relevance, it comes with challenges related to time, cost, and data reliability, which must be carefully managed to ensure a successful study.

3.3.3 Methods of Collecting Primary Data

Primary data can be gathered using various techniques, including observation, interviews, questionnaires, and schedules. Each method has its own advantages and is suitable for different research contexts. Below is a detailed examination of these methods.

(a) Observation Method

This approach involves collecting information firsthand through direct observation rather than relying on reports from others. The researcher records relevant details without posing direct questions to respondents, and in some cases, without their awareness. This technique is particularly useful in behavioral studies, such as analyzing

visitor behavior at trade fairs, assessing employee attitudes at work, or observing customer negotiation strategies. Observation can be classified into two types:

- **Participant Observation:** The researcher actively engages in the daily activities of individuals or organizations and observes their behavior from within.
- **Non-Participant Observation:** The researcher remains an external observer, monitoring activities without direct involvement.

Advantages

i) This method is ideal when respondents are unwilling or unable to share information.

ii) It provides deeper insights and generally yields accurate data that is quicker to process, making it more suitable for intensive studies rather than extensive ones.

Limitations

i) Events may occur unpredictably, making it difficult to ensure an observer is present at the right moment.

ii) Respondents may modify their behavior if they are aware of being observed.

iii) The observer's personal biases and lack of training can affect the accuracy of recorded observations.

iv) This method is not feasible for large-scale inquiries that cover a vast geographical area.

(b) Interview Method

In this method, the interviewer directly interacts with the respondents and gathers necessary information related to the research topic. Typically, the interviewer carries a structured set of questions or a questionnaire and follows it while conducting the interview. Through careful cross-examination, the interviewer ensures the accuracy and relevance of the collected data. To obtain reliable responses, the interviewer must possess strong communication skills and a strategic approach. Depending on the research requirements, interviews can be conducted in person (such as face-to-face or in-depth interviews) or over the phone. This method serves as a crucial tool in social research, as it involves a structured and direct conversation between the interviewer and the respondent. It facilitates the collection of pertinent information for a specific research issue. Interviews can be broadly categorized into direct personal interviews and indirect personal interviews.

Direct Personal Interviews

Under this method there is face to face contact between the interviewer and the interviewee. This sort of interview may be in the form of direct personal investigation or it may be indirect oral investigation. In the case of direct personal investigation the interviewer has to collect the information personally from the sources concerned. He has to be on the spot and has to meet people from whom data have to be collected. This method is particularly suitable for intensive investigations.

Indirect Personal Interviews (Telephone Interview)

Telephone interview involves trained interviewers phoning people to collect questionnaire data. This method is quicker and less expensive than face-to-face interviewing. However, only people with telephones can be interviewed, and the respondent can end the interview very easily.

Merits of the Interview Method

The major merits of this method are as follows:

i) People are generally more open to sharing information when approached directly, making personal interviews an effective way to achieve high response rates.

ii) This approach allows interviewers to clarify any uncertainties respondents may have, ensuring that the collected data is both reliable and valid.

iii) The interviewer can observe and analyze the respondent's reactions, gaining deeper insights into their responses.

iv) The interviewer can adapt the language and style of communication to suit the respondent's background, facilitating the collection of personal information that aids in better data interpretation.

Demerits of the Interview Method

Despite its advantages, the interview method has certain drawbacks:

i) There is a risk of bias, as the interviewer's personal opinions or perspectives may unintentionally influence the responses.

ii) Conducting interviews requires well-trained interviewers; otherwise, the accuracy and quality of data may be compromised.

iii) The method is often costly and time-consuming, especially when dealing with a large sample dispersed across different locations.

iv) It is not practical for large-scale research studies that require extensive data collection from a vast population.

(c) Through Local Reporters and Correspondents

Under this method, local investigators, agents, or correspondents are designated across various regions to gather information. Government agencies frequently employ this method for cases requiring continuous data collection. It is also widely used by newspapers, magazines, radio, and television news channels. This technique is suitable when consistent updates are needed, and absolute accuracy is not a primary concern.

Merits

i) This method is cheap and economical for extensive investigations.

ii) It gives results easily and promptly.

iii) It can cover a wide area under investigation.

Demerits

i) The data obtained may not be reliable.

ii) It gives approximate and rough results.

iii) It is unsuited where a high degree of accuracy is desired.

iv) As the agent/reporter or correspondent uses his own judgement, his personal bias may affect the accuracy of the information sent.

(d) Questionnaire and Schedule Methods

The questionnaire and schedule methods are widely used for gathering primary data in business research. Both approaches involve a structured set of questions arranged sequentially to facilitate the investigation. Let's explore these methods individually.

i) Questionnaire Method

Under this method, questionnaires are distributed to respondents either in person or via mail, requesting them to complete and return the form. When sent by post, it is referred to as a Mail Questionnaire. Depending on the nature of the study, time constraints, and available resources, questionnaires may also be sent through email. Upon receiving the questionnaire, respondents review the questions and provide their answers in the designated spaces. To enhance response rates, it is recommended to include a self-addressed envelope for easy return.

Merits of Questionnaire Method

i) This method can be used in cases where informants are spread over a vast geographical area.

ii) Respondents can take their own time to answer the questions. So the researcher can obtain original data by this method.

iii) This is a cheap method because its mailing cost is less than the cost of personal visits.

iv) This method is free from bias of the investigator as the information is given by the respondents themselves.

v) Large samples can be covered and thus the results can be more reliable and dependable.

Demerits of Questionnaire Method

i) Respondents may not return filled in questionnaires, or they can delay in replying to the questionnaires.

ii) This method is useful only when the respondents are educated and co-operative.

iii) Once the questionnaire has been dispatched, the investigator cannot modify the questionnaire.

iv) It cannot be ensured whether the respondents are truly representative.

ii) Schedule Method

A Schedule is also a list of questions, which is used to collect the data from the field. This is generally filled in by the researcher or the enumerators. If the scope of the study is wide, then the researcher appoints people who are called enumerators for the purpose of collecting the data. The enumerators go to the informants, ask them the questions from the schedule in the order they are listed and record the responses in the space meant for the answers in the schedule itself. For example, the population census all over the world is conducted through this method. The difference between questionnaire and schedule is that the former is filled in by the informants, the latter is filled in by the researcher or enumerator.

Merits of Schedule Method

i) It is a useful method in case the informants are illiterates.

ii) The researcher can overcome the problem of non-response as the enumerators go personally to obtain the information.

iii) It is very useful in extensive studies and can obtain more reliable data.

Demerits of Schedule Method

i) It is a very expensive and time-consuming method as enumerators are paid persons and also have to be trained.

ii) Since the enumerator is present, the respondents may not respond to some personal questions.

iii) Reliability depends upon the sincerity and commitment in data collection.

Difference between Schedule and Questionnaire

S. No	Questionnaire	Schedule
1.	A questionnaire is usually sent to respondents via mail along with a cover letter explaining the instructions, but no further assistance is provided by the sender.	A schedule is completed by a researcher or enumerator, who can clarify and interpret the questions if needed.
2.	Data collection through questionnaires is cost-effective, as expenses are mainly limited to designing and mailing the forms.	Data collection through schedules is more costly, as it requires hiring and training enumerators, along with preparing the schedules.
3.	The rate of non-response is generally high since many recipients either do not respond or return incomplete questionnaires. As a result, bias due to non-response remains uncertain.	The non-response rate is low because enumerators ensure that all questions are answered. However, there is still a risk of interviewer bias and potential manipulation.

4.	The identity of the respondent remains uncertain.	The respondent's identity is known.
5.	The process of collecting data through questionnaires is typically slow, as many recipients delay or fail to return them.	Data collection is timely and efficient since schedules are completed by enumerators.
6.	There is no personal interaction, as questionnaires are mailed and returned through postal services.	Direct personal interaction occurs between the enumerator and the respondent.
7.	A broader and more representative sample can be covered, as questionnaires can be distributed widely.	Expanding the sample size is challenging due to the logistical difficulties of deploying enumerators across large areas.

Self-Check Exercise 3.1

Q1. What is meant by Primary Data.

Q2. Explain the merits and demerits of Primary Data.

Q3. List out the various methods of collecting Primary Data.

Q4. What is observation method.

Q5. Distinguish between Questionnaire and Schedule.

3.4 SECONDARY DATA

Secondary data is the data that has already been searched by the researchers for their purpose and people can access these gathered resources through different journals, books, websites, etc. These sources are available instantly as compared to primary data. The time and effort required to collect data are less. Secondary data by all means are an effective and efficient way to analyze data. But there are a lot of other factors that may lead to misrepresentation or misleading data as the data received may not be relevant for the study. Therefore there are many advantages and disadvantages of secondary data. Let us see the advantages and disadvantages.

3.4.1 Advantages of Secondary Data

The following are the advantages of secondary data:

i) **Easy to access**: Data is available anywhere and anytime it can be in the form of periodicals, magazines, or can be accessed anytime through the internet. People generally use secondary data maximum nowadays to evaluate their studies. A very small example is the students who depend on books, internet sites, and teachers to access information and prepare for exams.

ii) **Low cost or cost-effective**: The secondary data is of low cost as data are available at cheap rates, for example, the internet access, newspaper, or periodicals are available at cheaper rates and available in large quantities, so there is no non-availability of data to its users. Thus it is cost-effective.

iii) **Less time taking**: Data is available quickly and readily while primary data need to be collected first and then only after summarization data are used. Time taken to collect and analyze data is less than secondary data that is quickly available. Therefore it takes less time to take the source of data.

iv) **Various sources are available to collect data:** Secondary data is not only available through one source, but there are multiple sources like books, magazines, the internet, periodicals, and many more. Therefore various sources are available to collect data for analysis for its users. These sources are easily accessible and readily available to their users.

v) **Data can be collected by anyone**: Anyone can collect data whether he /she is specialized in collecting it or not, depending upon the use. Also, there is no ownership of data that can be claimed by its user as data has already been shared by its owner, who was a primary collector of data.

vi) **The study is based on longitudinal analysis:** Since the data has been collected over years, thus a longitudinal analysis is done by the researchers with the help of secondary data. The data collected is more reliable and valid for users.

3.4.2 Disadvantages of Secondary Data

i) **Inaccuracy**: It is a limitation of secondary data that the data collected over the past few years may be inaccurate. The basis of data collected may not be correct or the analysis or interpretation made may not be accurate or relevant.

ii) **Data may be sometimes outdated:** The data provided through different sources may also be outdated as it has been stored and managed for many years. Therefore it may also sometimes be outdated and may not be relevant for today's scenario.

iii) **Not compatible with the needs of the user:** Since data is related to past surveys and according to the needs of the researchers of that time. It may happen that the present user of this data may not need or not have topics relevant to his study or research. Therefore here instead of outdated data, the data becomes irrelevant for the user to be used in research.

iv) **Anyone can access data**: There is no privatization of data by its owner, data can be accessed by anyone willing to research on that topic. There is no secrecy of data but the user of data cannot appeal their possession or ownership of the data they accessed.

v) **Data quality cannot be controlled:** The researchers have no control over the quality of data. As data is already surveyed by researchers according to their relevant basis and there may be changes in the surroundings and other factors that may lead to the change in the data provided thus no proper quality can be controlled.

vi) **Data can be biased**: Since data collected by the researcher is based on his/her opinion, therefore data is biased. And it may also have an impact on the data collected by the user of the secondary data.

3.4.3 Sources of Secondary Data

The sources of secondary data are classified into the following two categories:

(1) Documentary Sources of Data

This category of secondary data source may also be termed as Paper Source. The main sources of documentary data can be broadly classified into two categories:

a) Published sources, and

b) Unpublished sources.

a) Published Sources

There are various national and international institutions, semi-official reports of various committees and commissions and private publications which collect and publish statistical data relating to industry, trade, commerce, health etc. These publications of various organisations are useful sources of secondary data. These are as follows:

i) **Government Publications:** Central and State Governments publish current information alongwith statistical data on various subjects, quarterly and annually. For example, Monthly Statistical Abstract, National Income Statistics, Economic Survey, Reports of National Council of Applied Economic Research (NCEAR), Federation of Indian Chambers of Commerce and Industry (FICCI), Indian Council of Agricultural Research (ICAR), Central Statistical Organisation (CSO), etc.

ii) **International Publications:** The United Nations Organisation (UNO), International Labour Organisation (ILO), International Monetary Fund (IMF), World Bank, Asian Development Bank (ADB) etc., also publish relevant data and reports.

iii) **Semi-official Publications:** Semi-official organisations like Corporations, District Boards, Panchayat etc. publish reports.

iv) **Committees and Commissions:** Several committees and commissions appointed by State and Central Governments provide useful secondary data. For example, the report of the 10th Financial Commission or Fifth Pay Commissions etc.

v) **Private Publications:** Newspapers and journals publish the data on different fields of Economics, Commerce and Trade. For example, Economic Times, Financial Express etc. and Journals like Economist, Economic and Political Weekly, Indian Journal of Commerce, Journal of Industry and Trade, Business Today etc. Some of the research and financial institutions also publish their reports annually like Indian Institute of Finance. In addition to this, reports prepared by research scholars, universities etc. also provide secondary source of information.

b) Unpublished Sources

It is not necessary that all the information/data maintained by the institutions or individuals are available in published form. Certain research institutions, trade associations, universities, research scholars, private firms, business institutions etc., do collect data but they normally do not publish it. We can get this information from their registers, files etc.

(2) Electronic Sources

The secondary data is also available through electronic media (through Internet). Data can be downloaded from web sites like google.com; yahoo.com; etc., and typing your subject for which the information is needed.

3.3.1 Precaution in using Secondary Data

With the above discussion, we can understand that there is a lot of published and unpublished sources where researcher can gets secondary data. However, the researcher must be cautious in using this type of data. The reason is that such type of data may be full of errors because of bias, inadequate size of the sample, errors of definitions etc. Hence, before using secondary data, you must examine the following points.

i) Suitability of Secondary Data

Before using secondary data, you must ensure that the data are suitable for the purpose of your enquiry. For this, you should compare the objectives, nature and scope of the given enquiry with the original investigation.

ii) **Reliability of Secondary Data**

For the reliability of secondary data, these can be tested: i) unbiasedness of the collecting person, ii) proper check on the accuracy of field work, iii) the editing, tabulating and analysis done carefully, iv) the reliability of the source of information, v) the methods used for the collection and analysis of the data. If the data collecting organisations are government, semi-government and international, the secondary data are more reliable corresponding to data collected by individual and private organisations.

iii) Adequacy of Secondary Data

Adequacy of secondary data is to be judged in the light of the objectives of the research. Adequacy of the data may also be considered in the light of duration of time for which the data is available.

Self-Check Exercise 3.2

Q1. What is meant by Secondary Data?

Q2. Explain the merits and demerits of Secondary Data.

Q3. List out the various sources of Secondary Data.

Q4. What precaution do we need in using Secondary Data.

3.5 SELECTION OF APPROPRIATE METHOD FOR DATA COLLECTION

Given the wide range of data collection methods available, selecting the most suitable one depends on the specific research context. The following factors should be considered when making this decision:

- **Nature, Scope, and Objectives of the Study:** The chosen method should align with the study's purpose, scope, and objectives. Researchers must determine

whether existing data (secondary data) can be utilized or if new data (primary data) need to be collected.

- **Availability of Funds:** The financial resources available play a crucial role in method selection. If a method is too costly, it may be impractical to implement within the given budget constraints.
- **Time Considerations:** The time available for data collection significantly influences method selection. Some methods require extended periods, while others allow for quicker data gathering. Researchers must choose an approach that fits within the allotted timeframe.
- **Required Level of Accuracy:** The degree of precision needed in the study is another essential factor. Some methods yield highly accurate data, while others provide estimates that may be sufficient depending on the research goals.

Self-Check Exercise 3.3

Q1. What factors affect an appropriate method for data collection?

3.6 SUMMARY

The reliability of research results depends on the quality of data. The quality of data can be expressed in terms of its representative feature of the reality which can be ensured by the usage of fitting data collection methods. Statistical data can be classified in two categories, viz. primary and secondary. Primary data is a type of information that is obtained directly from first-hand information by means of surveys, observations of experimentation. Primary data can be collected by observation method, questionnaire methods, interview method and interview schedule method. On the other hand, secondary data is the data that have been already collected by the others and readily available from other sources, however, researcher must be very careful in using secondary data. Thus, before using secondary data, researcher must check the reliability, suitability and adequacy of such data. As there are many methods for collection of data, it is important to consider the factors such as nature scope and object of enquiry, availability of funds, time factor, and degree of accuracy in mind while selecting a particular method.

3.7 GLOSSARY

- **Data:** Data is the raw material from which useful information is derived.
- **Primary data** are always collected from the source. It is collected either by the investigator himself or through his agents.
- **Secondary data** are those which have already been collected by someone and have gone through the statistical machines. They are usually refined of the raw materials.
- **Questionnaires**: are sent personally or by post to various informants with a request to answer the questions and return the questionnaire.
- **Schedule**: is also a list of questions, which is used to collect the data from the field. This is generally filled in by the researcher or the enumerators

3.8 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 3.1

Ans. Q1. Refer to Section 3.3
Ans. Q2. Refer to Sections 3.3.1 and 3.3.2
Ans. Q3. Refer to Section 3.3.3
Ans. Q4. Refer to Section 3.3.3 (a)
Ans. Q5. Refer to Section 3.3.3 (d)

Self-Check Exercise 3.2

Ans. Q1. Refer to Section 3.4
Ans. Q2. Refer to Sections 3.4.1 and 3.4.1
Ans. Q3. Refer to Section 3.4.3
Ans. Q4. Refer to Section 3.4.4

Self-Check Exercise 3.3

Ans. Q1. Refer to Section 3.5

3.9 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House.

3.10 TERMINAL QUESTIONS

Q1. What is primary data? Discuss the various methods of collecting primary data?

Q2. What are the merits and demerits of secondary data? Explain various sources of secondarydata.

Q3. What do you mean by questionnaire? What is the difference between questionnaire andschedule?

Unit-04

GRAPHIC PRESENTATION OF DATA

STRUCTURE

4.1 INTRODUCTION

In Unit-02, we introduced you to the data and its different types. This unit deals with the graphical presentation of the data. The most commonly used graphs and curves for representation a frequency distribution are Histogram, Frequency Polygon, Smoothened Frequency Curve and Ogives or Cumulative Frequency Curves. In this unit, we will study in detail each of them.

4.2 LEARNING OBJECTIVES

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

- Explain the Histograms
- Elucidate frequency Polygon
- Describes Ogives
- Explain bar diagram and Pie Charts

4.3 GRAPHS OF FREQUENCY DISTRIBUTION

A frequency distribution can be represented graphically in any of the following ways. The most commonly used graphs and curves for representation a frequency distribution are

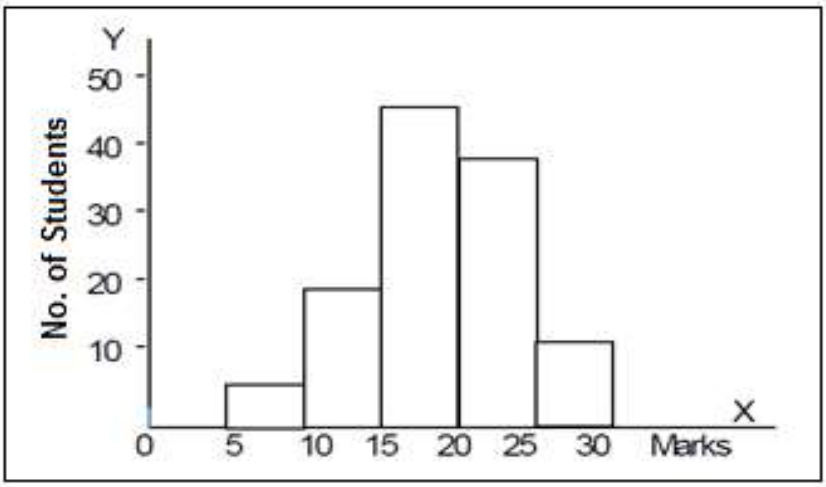
4.3.1 Histogram
4.3.2 Frequency Polygon
4.3.3 Smoothened Frequency Curve
4.3.4 Ogives or Cumulative Frequency Curves.
4.3.5 Bivariate Frequency Distribution

4.3.1 Histogram

A histogram is a set of vertical bars whose one as are proportional to the frequencies represented. While constructing histogram, the variable is always taken on the X axis and the frequencies on the Y axis. The width of the bars in the histogram will be proportional to the class interval. The bars are drawn without leaving space between them. A histogram generally represents a continuous curve. If the class intervals are uniform for a frequency distribution, then the width of all the bars will by equal.

Example 1: Draw a histogram to represent the following data

Marks	10-15	15-20	20-25	25-30	30-35
No. of Students	5	20	47	38	10



4.3.2 Frequency Polygon (or Line Graphs)

Frequency Polygon is a graph of frequency distribution. There are two ways of constructing a frequency polygon.

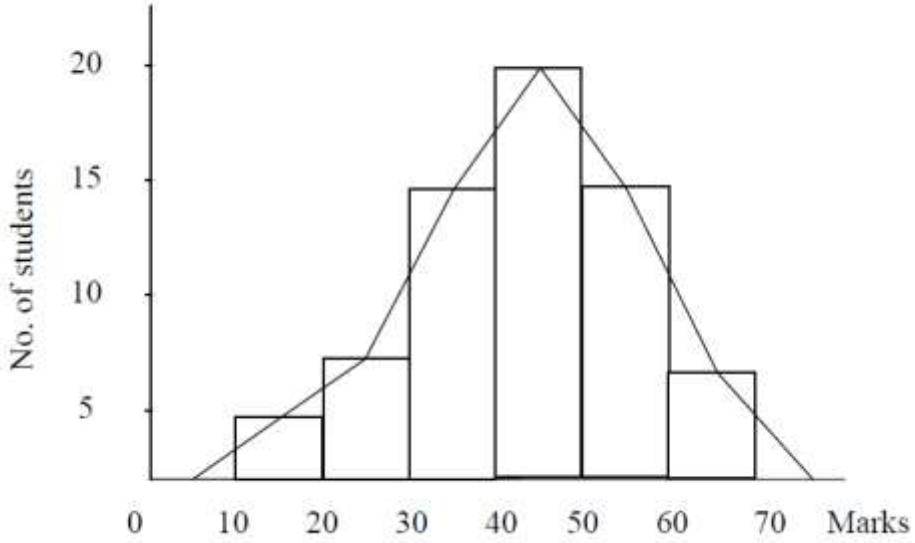
a) Draw histogram of the data and then join by straight lines the mid points of upper horizontal sides of the bars. Join both ends of frequency polygon with x axis. Then we get frequency polygon.

b) Another method of constructing frequency polygon is to take the mid points of the various class intervals and then plot frequency corresponding to each point and to join all these points by a straight line.

Here we have not to construct a histogram:-

Example 2: Draw a frequency polygon to the following frequency distribution

Marks	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	5	8	15	20	12	7



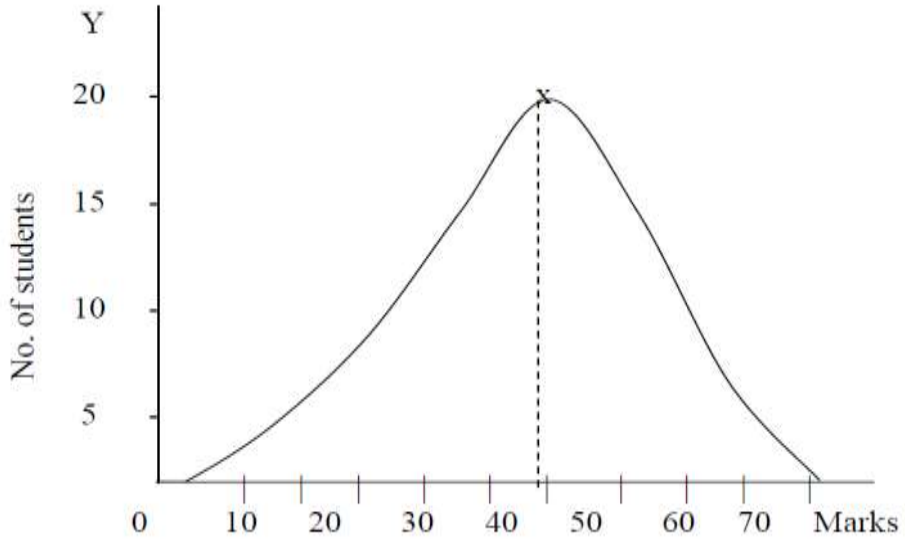
4.3.3 Frequency Curves

A continuous frequency distribution can be represented by a smoothed curve known as frequency curve. The mid values of classes are taken along the x axis and the frequencies along y axis. The points thus plotted are joined by smoothened curve. When the points of a frequency polygon are joined by free hand method curve and not by a straight line, we get frequency curve.

The curve is drawn freehand in such a manner that the area included under the curve is approximately same as that of the frequency polygon. If the class intervals are not uniform, adjust they y co-ordinate so that the frequencies are proportional to the area of the rectangle contained by plotted points.

Example 3: Draw a frequency curve to the following frequency distribution

Marks	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	5	8	15	20	12	7



4.3.3.1 Difference between frequency polygon and frequency curve

Frequency polygon is drawn to frequency distribution of discrete or continuous nature. Frequency curves are drawn to continuous frequency distribution. Frequency polygon is obtained by joining the plotted points by straight lines. Frequency curves are smooth. They are obtained by joining plotted points by smooth curve.

4.3.4 Ogives or Cumulative Frequency Curve

A frequency distribution when cumulated, we get cumulative frequency distribution. A series can be cumulated in two ways. One method is frequencies of all the preceding classes' one added to the frequency of the classes. This series is called less than cumulative series. Another method is frequencies of succeeding classes are added to the frequency of a class. This is called more than cumulative series. Smoothed frequency curves drawn for these two cumulative series are called cumulative frequency curve or Ogives. Thus corresponding to the two cumulative series we get two ogive curves, known as less than ogive and more than ogive.

4.3.4.1 Less than Ogive Curve is obtained by plotting frequencies (cumulated) against the upper limits of class intervals.

4.3.4.2 More than Ogive Curve is obtained by plotting cumulated frequencies against the lower limits of class intervals. Less than ogive is an increasing curve, slopping upwards from left to right. More than ogive is a decreasing curve and slopes from left to right.

Example 4: Draw less than and more than cumulative frequency distribution for the following frequency distribution.

Marks	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	4	6	10	20	18	2

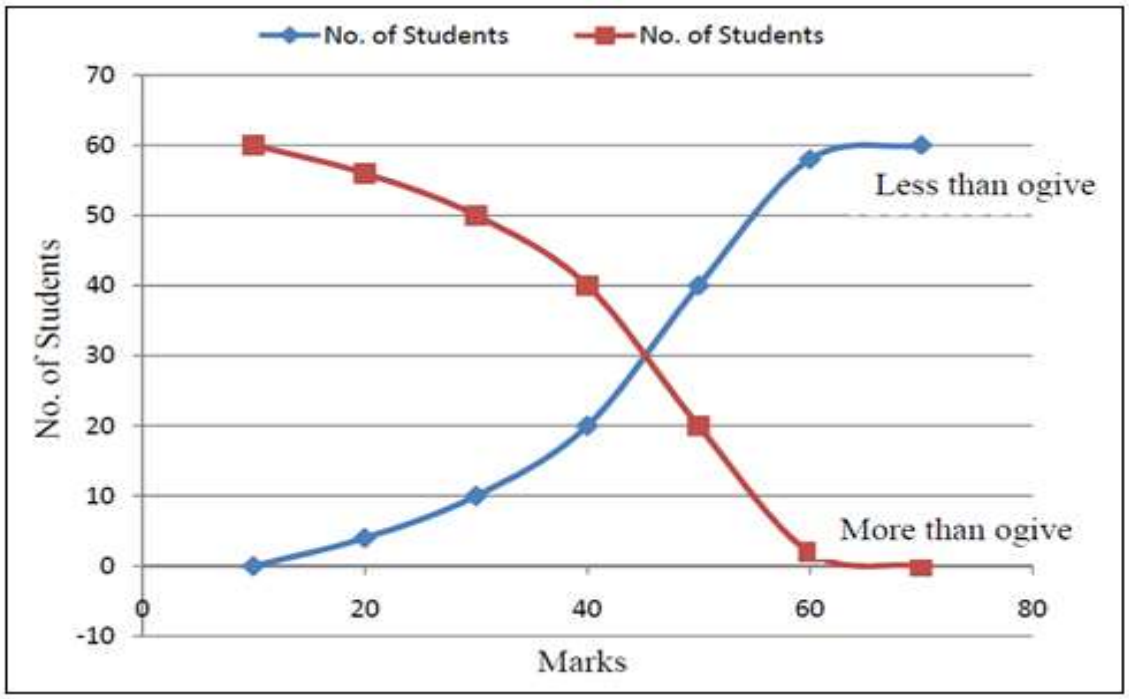
Solution

a) Less than Cumulative frequency distribution**:**

Marks	**No. of Students**	**Marks less than**	**No. of Students**
0-10	0	10	0
10-20	4	20	4
20-30	6	30	10
30-40	10	40	20
40-50	20	50	40
50-60	18	60	58
60-70	2	70	60

b) More than Cumulative frequency distribution:

Marks	**No. of Students**	**Marks More than**	**No. of Students**
0-10	0	10	60
10-20	4	20	56
20-30	6	30	50
30-40	10	40	40
40-50	20	50	20
50-60	18	60	2
60-70	2	70	0



4.3.5 Bivariate Frequency Distribution

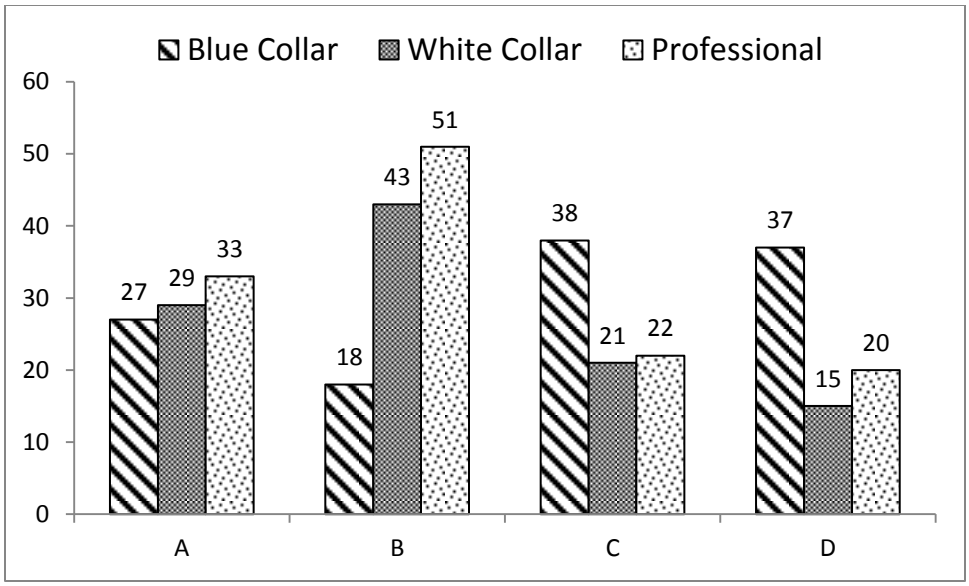
In case the data involve two variables (such as profit and expenditure on advertisements of a group of companies, income and expenditure of a group of individuals, supply and demand of a commodity, etc.), then frequency distribution so obtained as a result of cross classification is called bivariate frequency distribution. It can be summarized in the form of a two-way (bivariate) frequency table and the values of each variable are grouped into various classes (not necessarily same for each variable) in the same way as for univariate distributions.

Describing the Relationship between Two Nominal Variables (Bivariate)

Bivariate analysis shows the relationship between two variables. The following tables are to be used for nominal data.

Cross Tabulation/Classification Table

Occupation	**Types of Company**				**Total**
	A	**B**	**C**	**D**	
Blue Collar	27	18	38	37	120
White Collar	29	43	21	15	108
Professional	33	51	22	20	126
Total	**89**	**112**	**81**	**72**	**354**



The above table and diagram sho the frequency of data fitting 2 variables. For example, 33 people work for G&M and are 'professionals'. To get the totals on the right hand side, we simply sum up each column number of the same row. For example 27 + 18 + 38 + 38 = 120.

Self-Check Exercise 4.1

Q1. What is meant by histogram

Q2. Draw a frequency polygon to the following frequency distribution

Marks	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	5	8	15	20	12	7

Q3. Difference between frequency polygon and frequency curve

4.4 PIE DIAGRAMS

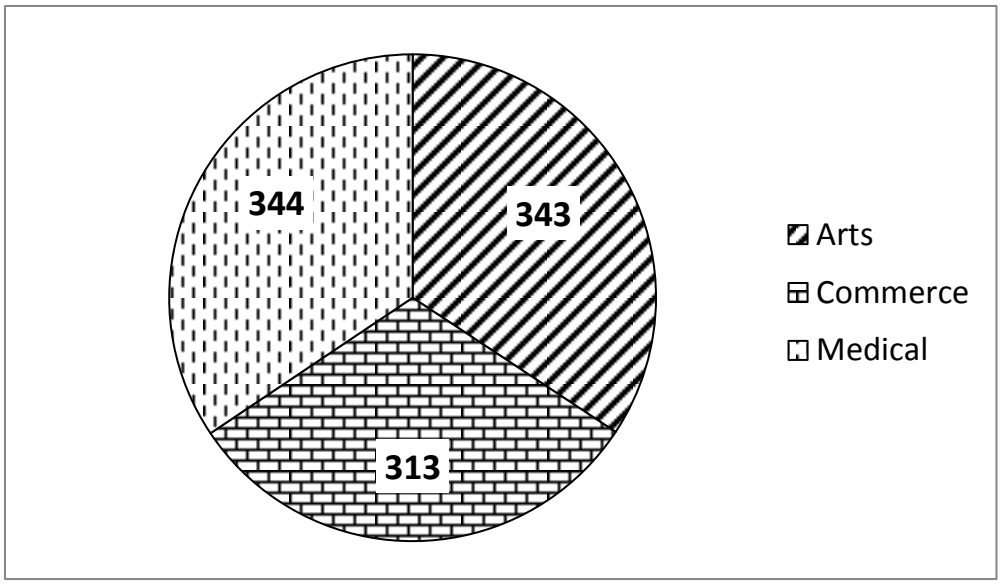
Pie diagrams are used when the aggregate and their division are to be shown together. The aggregate is shown by means of a circle and the division by the sectors of the circle. For example to show the total expenditure of a government distributed over different departments like agriculture, irrigation, industry, transport etc. can be shown in a pie diagram. In constructing a pie diagram the various components are first expressed as a percentage and then the percentage is multiplied by 3.6. So we get angle for each component. Then the circle is divided into sectors such that angles of the components and angles of the sectors are equal. Therefore one sector represents one component. Usually components are with the angles in descending order are shown.

Example 5: Draw pie diagram to represents the distribution of the students among Arts, Commerce, and Medical courses.

Course	Arts	Commerce	Medical	Total
Number of Students	343	313	344	1000

Solution:

Race	**Number of Students**	**% age**	**Angle (% x 3.6)**
Arts	343	34.3	123.48
Commerce	313	31.3	112.68
Medical	344	34.4	123.84
	1000	100.00	360.00



Self-Check Exercise 4.2

Q1. What are pie diagram

Q2: Draw pie diagram to represents the distribution of the students among Arts, Commerce, and Medical courses.

Course	Arts	Commerce	Medical	**Total**
Number of Students	343	313	344	**1000**

4.5 BAR GRAPHS

Bar graphs serve as an effective tool for visually representing data, similar to line graphs. However, instead of using plotted points to indicate values, bar graphs utilize rectangular bars, which can be oriented either vertically or horizontally. These bars extend to a specific length that corresponds to the data being represented. There are several key advantages that make bar graphs particularly useful:

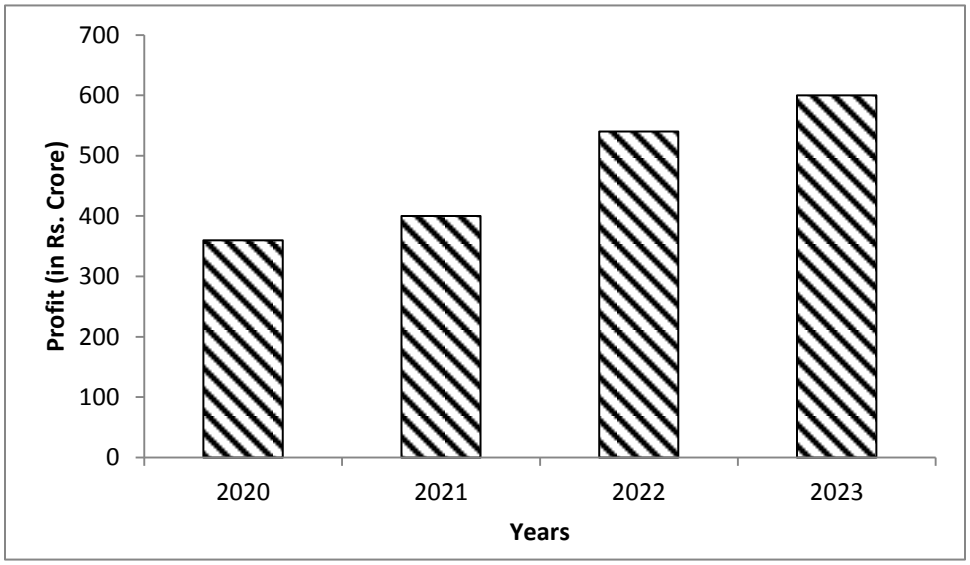
i) They allow for easy comparison between different variables.

ii) They effectively illustrate trends in data, showing how one variable changes in response to another.

iii) They enable quick determination of one variable's value when the other is known.

Example 6: Draw bar diagram to represents the year wise profits of XYZ Company.

Year	2020	2021	2022	2023
Profit (in Rs. Crore)	360	400	540	600



Self-Check Exercise 4.3

Q1. What are Bar Graphs

4.6 SUMMARY

In this unit, we have studied about the graphical presentation of data**.** The most commonly used graphs and curves for representing a frequency distribution are histogram, frequency polygon, frequency curves and ogives- all these we have studied in this unit.

4.7 GLOSSARY

- **Histogram:** A histogram is a set of vertical bars whose one as are proportional to the frequencies represented. While constructing histogram, the variable is always taken on the X axis and the frequencies on the Y axis. The width of the bars in the histogram will be proportional to the class interval.
- **Frequency Polygon** is a graph of frequency distribution.
- **Frequency Curves**: A continuous frequency distribution can be represented by a smoothed curve known as frequency curve. The mid values of classes are taken along the x axis and the frequencies along y axis. The points thus plotted are joined by smoothened curve. When the points of a frequency polygon are joined by free hand method curve and not by a straight line, we get frequency curve.

- **Ogives (Cumulative frequency curve)**: A frequency distribution when cumulated, we get cumulative frequency distribution.
- **Less than ogive curve** is obtained by plotting frequencies (cumulated) against the upper limits of class intervals.
- **More than ogive curve** is obtained by plotting cumulated frequencies against the lower limits of class intervals.

4.8 ANSWERS TO SELF-CHECK EXERCISES

Self-Check Exercise 4.1

Ans. Q1. Refer to Section 4.3

Ans. Q2. Refer to Section 4.3 (Example 2)

Ans. Q3. Refer to Section 4.3.3.1

Self-Check Exercise 4.2

Ans. Q1. Refer to Section 4.4

Ans. Q1. Refer to Section 4.4 (Example 5)

Self-Check Exercise 4.3

Ans. Q1. Refer to Section 4.5

4.9 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House.

4.10 TERMINAL QUESTIONS

Q1. What are ogives? How is less than ogive curve different from more than ogive curve?

Q2. What are frequency curve and frequency polygen?

Unit-05

APPLICATIONS OF COMPUTER IN GRAPHIC PRESENTATION

STRUCTURE

5.1 INTRODUCTION

Computers have become a powerful tool for the rapid and economical production of pictures. There is virtually no area in which graphical displays cannot be used to some advantage, and so it is not surprising to find the use of computer graphics so

widespread. The computer is an information processing machine. It is a tool for storing, manipulating and correlating data. There are many ways to communicate the processed information to the user. The computer graphics is one of the most effective and commonly used ways to communicate the processed information to the user. It displays the information in the form of graphics objects such as pictures, charts, graphs and diagrams instead of simple text. Thus we can say that computer graphics makes it possible to express data in pictorial form. The picture or graphics object may be an engineering drawing, business graphs, architectural structures, a single frame from an animated movie or a machine parts illustrated for a service manual.

5.2 LEARNING OBJECTIVES

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

- Explain computer graphics
- Explain basic issues in designing and implementing graphic presentations
- Explain the types of graphics presentation

5.3 COMPUTER GRAPHICS

In computer graphics, pictures or graphics objects are presented as a collection of discrete picture elements called pixels. The pixel is the smallest addressable screen element. It is the smallest piece of the display screen which we can control. The control is achieved by setting the intensity and color of the pixel which compose the screen. Although early applications in engineering and science had to rely on expensive and cumbersome equipment, advances in computer technology have made interactive computer graphics a practical tool. Today, we find computer graphics used routinely in such diverse areas as science, engineering, medicine, business**,** industry, government, art, entertainment, advertising, education, and training. Before we get into the details of how to do computer graphics, we first take a short tour through a gallery of graphics applications.

5.3.1 Advantages of Computer Graphics

i) A high quality graphics displays of personal computer provide one of the most natural means of communicating with a computer.

ii) It has an ability to show moving pictures, and thus it is possible to produce animations with computer graphics.

iii) With computer graphics use can also control the animation by adjusting the speed, the portion of the total scene in view, the geometric relationship of the objects in the scene to one another, the amount of detail shown and so on.

iv) The computer graphics also provides facility called update dynamics. With update dynamics it is possible to change the shape, color or other properties of the objects being viewed.

v) With the recent development of digital signal processing (DSP) and audio synthesis chip the interactive graphics can now provide audio feedback along with the graphical feedbacks to make the simulated environment even more realistic.

Self-Check Exercise 5.1

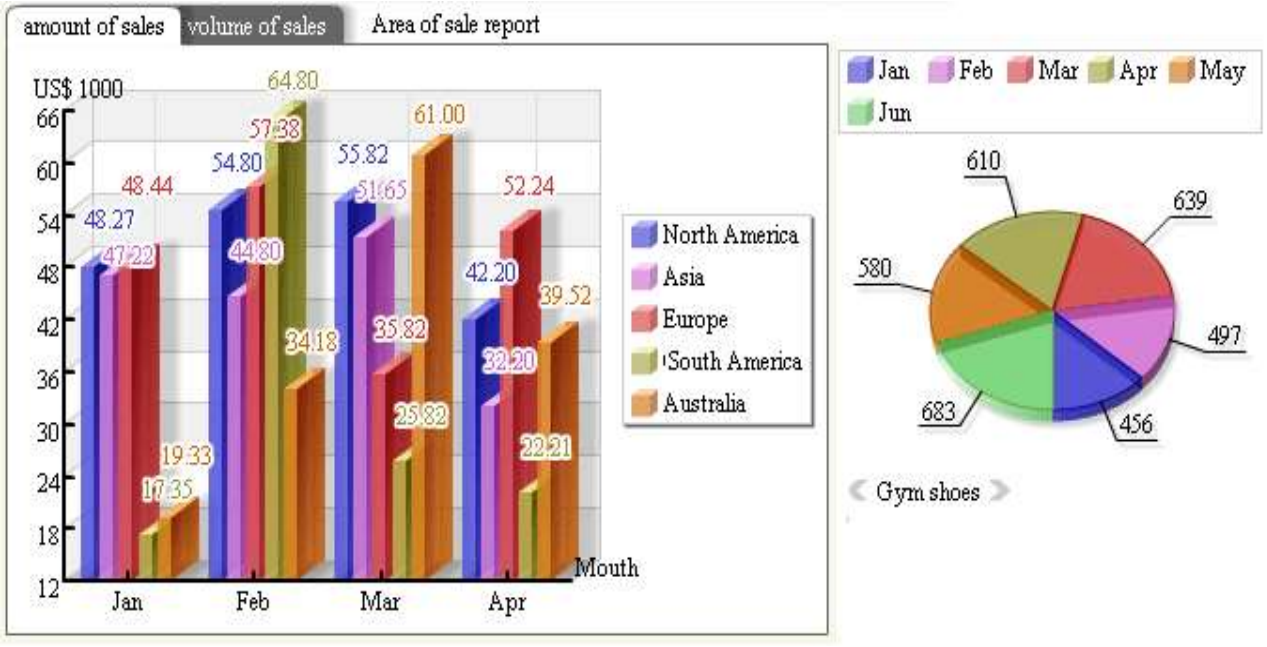
Q1. What do you understand by Computer graphics?

Q2. What are the advantages of computer graphics?

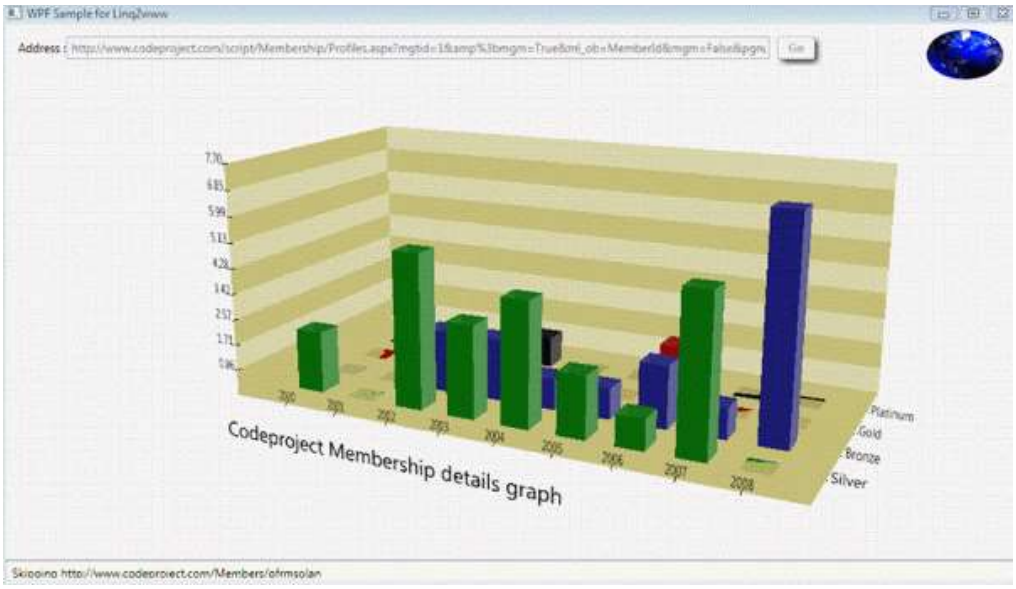
5.4 PRESENTATION GRAPHICS

One of the major application areas is presentation graphics, used to produce illustrations for reports or to generate 35-mm slides or transparencies for use with projectors. It is used to produce illustrations for reports or to generate slide for with projections. Examples of presentation graphics are bar charts, line graphs, surface graphs, pie charts and displays showing relationships between parameters. 3-D graphics can provide more attraction to the presentation. Presentation graphics is commonly used to summarize financial, statistical, mathematical, scientific, and economic data for research reports, managerial reports, consumer information bulletins, and other types of reports.

Workstation devices and service bureaus exist for converting screen displays into 35-mm slides or overhead transparencies for use in presentations. Typical examples of presentation graphics are bar charts, line graphs, surface graphs, pie charts, and other displays showing relationships between multiple parameters.



The above Figure gives examples of two-dimensional graphics combined with the sale of shoes in 5 different continents for first 6 months. This illustration shows 5 color coded bar charts combined on to one graph and a pie chart with six sections.



The above figure- 3D bar chart for code project membership details similar graphs and charts can be displayed in three dimensions to provide additional information. Three-dimensional graphs are sometime used simply for effect; they can provide a more dramatic or more attractive presentation of data relationships.

The charts displayed in the figure above include a three-dimensional bar graph. Presentation graphics serve three primary functions:

i) An editing tool that enables users to insert, modify, and format text.

ii) A feature for adding and manipulating graphic images.

iii) A slide show system designed to present content effectively.

Presentation graphics software is a specialized tool that assists users in creating visually engaging materials such as visual aids, handouts, and overhead slides. It allows for the integration of artwork, graphics, and text to develop a sequence of slides, aiding speakers in effectively delivering their messages.

It is important to note that presentation graphics encompass more than just PowerPoint presentations. They include various forms of visual representation, such as slide presentations, bar charts, pie charts, graphs, and multimedia presentations. The main advantage of using such software is that it facilitates the visual representation of abstract ideas, making information more comprehensible and impactful.

Self-Check Exercise 5.2

Q1. What do you understand by Presentation Graphics?

5.5 TYPES OF PRESENTATION GRAPHICS

5.5.1 Charts

1. Charts are used to help people understand numerical data through visualization

2. Appropriate charts can provide different perspectives, details, overviews, generalizations, and trends in data.
3. These visual language devices filter knowledge and provide appropriate chunking, structuring, and pacing in the presentation of data.
4. Scientists often distinguish charts from graphs.
5. Charts present numerical relations on a comparative basis.
6. Graphs present a functional relation between dependent and independent variables.

a) Bar chart,
b) pie chart,
c) various pictorial charts
d) Continuous line plots on linear-linear coordinate scales,
e) on semi-logarithmic coordinate scales,
f) on log-log coordinate scales

Basic issues in designing and implementing graphic presentations:

1. Simplicity, clarity, and consistency are essential for good chart design.
2. Keep extraneous text to a minimum.
3. Both legibility and readability can be significantly improved through the selection of graphic elements and the layout of the material.
4. Legibility deals with the reader's ability to successfully find, identify, and absorb what a chart denotes
5. Readability concerns the chart's interpretation and appeal.

Spreadsheet packages offer a variety of chart types for accomplishing this.

Self-Check Exercise 5.3

Q1. What are the types of Presentation Graphics?

5.6 EXCEL SUPPLIES

1. Column
2. Bar
3. Line
4. Pie
5. XY (scatter)
6. Area
7. Doughnut

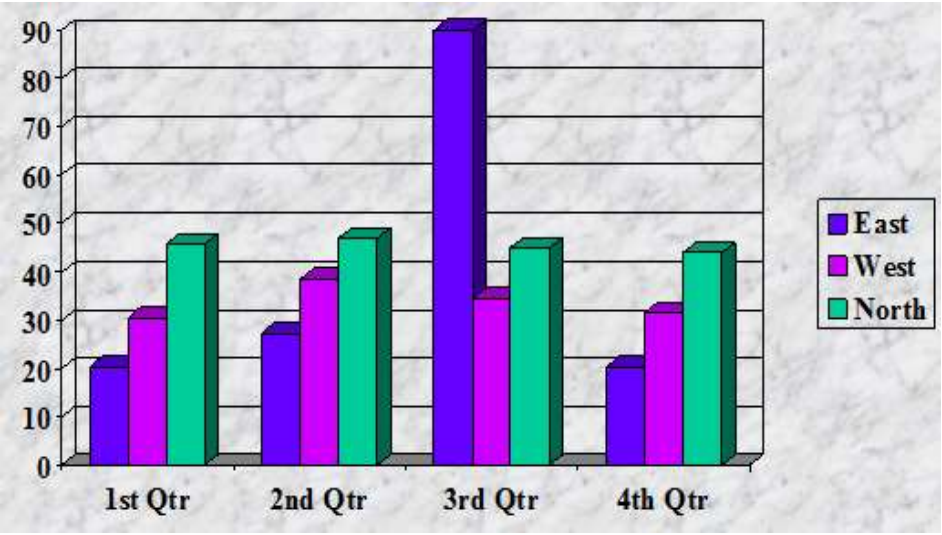
8. Radar
9. Surface
10. Bubble
11. Stock
12. Cone, cylinder, and pyramid

- You can create any one of these in Excel by first selecting the cells whose values are to be displayed in the chart.
- Then click the Chart Wizard button on the standard toolbar and choose from among the options presented on the resulting series of dialog boxes

5.6.1 Column Charts

1) A column chart shows the data as vertical bars.
2) A column chart shows data changes over a period of time or illustrates comparisons among items.
3) Categories are organized horizontally, values vertically, to emphasize variation over time

a. Stacked column charts show the relationship of individual items to the whole.
b. Step charts are column charts without space between the columns.
c. The 3-D perspective column chart compares data points along two axes. In this 3-D chart, you can compare four quarters of sales performance in one division with the performance of two other divisions.

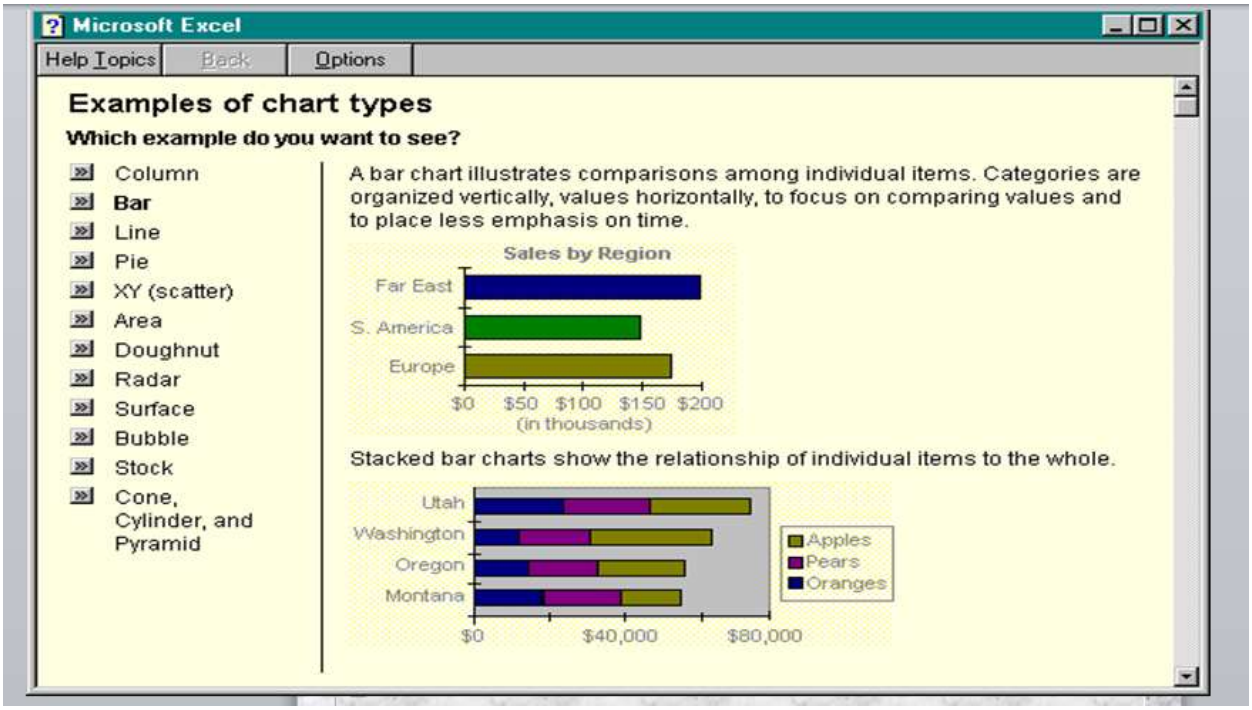


5.6.2 Bar Chart

1) A bar chart displays the data as horizontal bars.
2) It is used to illustrate comparisons among individual items.

3) Categories are organized vertically, values horizontally, to focus on comparing values and to place less emphasis on time.

4) Stacked bar charts show the relationship of individual items to the whole.



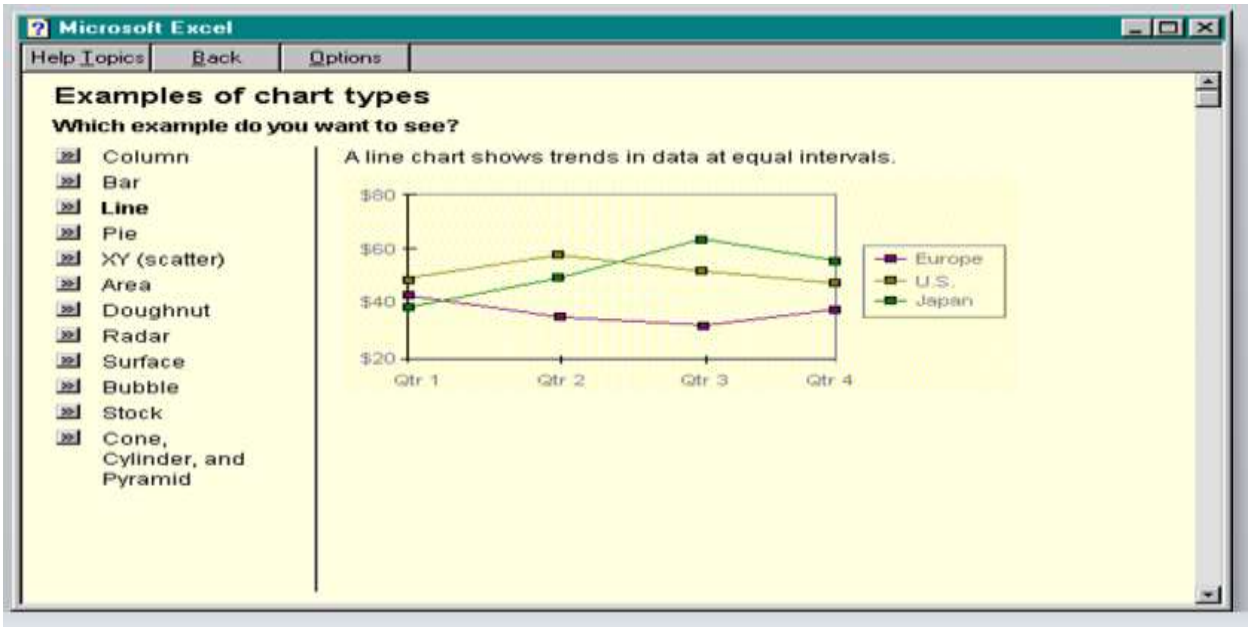
5.6.3 Line chart

1) A line chart shows data plotted with or without markers at the data point locations and with or without line segments connecting the points.

2) A line chart shows trends in data at equal intervals, discretely or continuously.

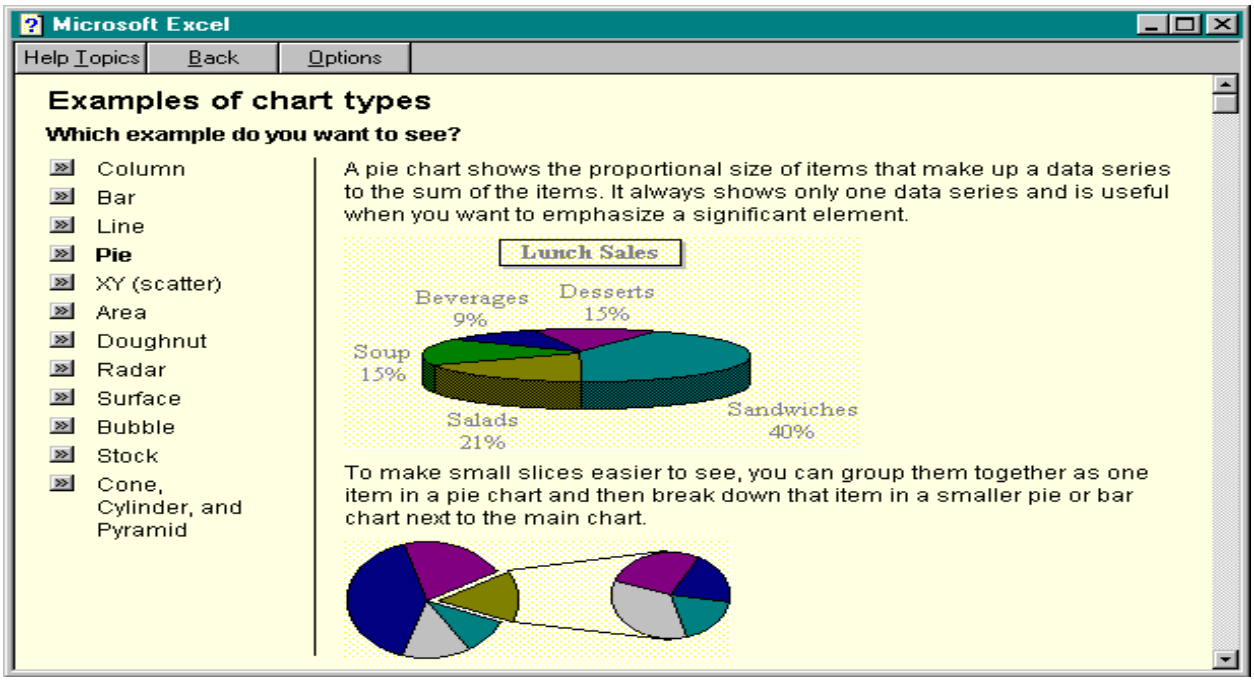
3) A line chart shows data plotted with or without markers at the data point locations and with or without line segments connecting the points.

4) A line chart shows trends in data at equal intervals, discretely or continuously.



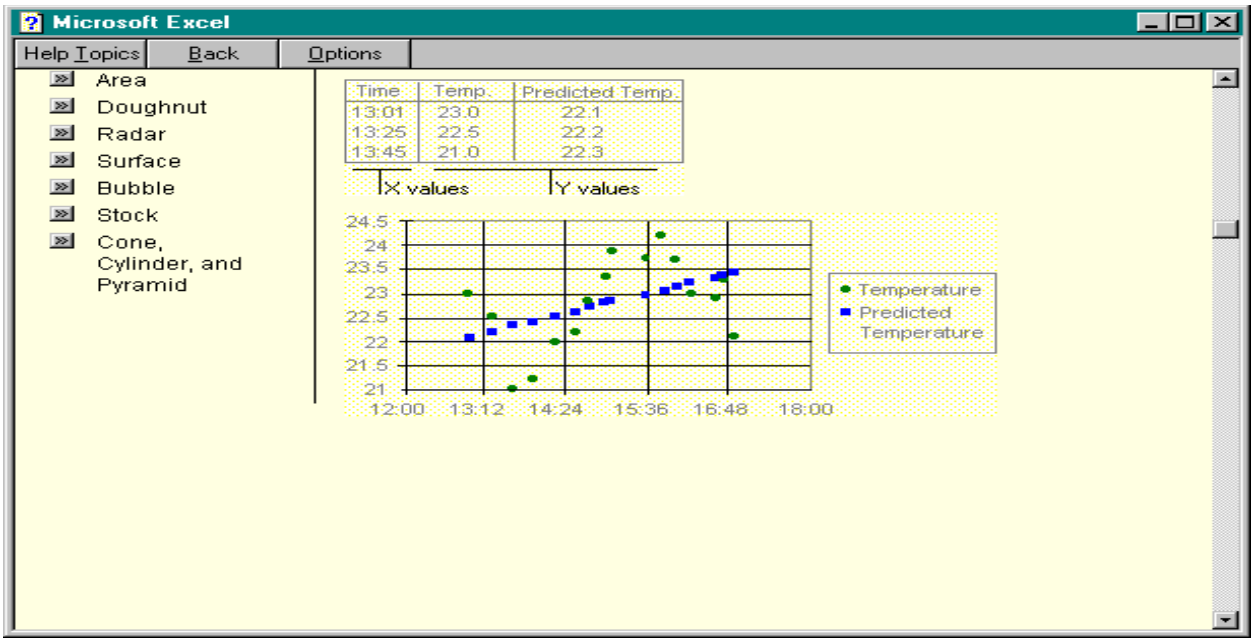
5.6.4 Pie Chart

1) A pie chart displays the data in a circle divided into sectors.
2) It shows the proportional size of items (out of 100%) that make up a data series to the sum of the items.
3) It always shows only one data series and is useful when you want to emphasize a significant element.
4) To make small sectors/slices easier to see, you can group them together as one item in a pie chart and then break down that item in a smaller pie or bar chart next to the main chart.



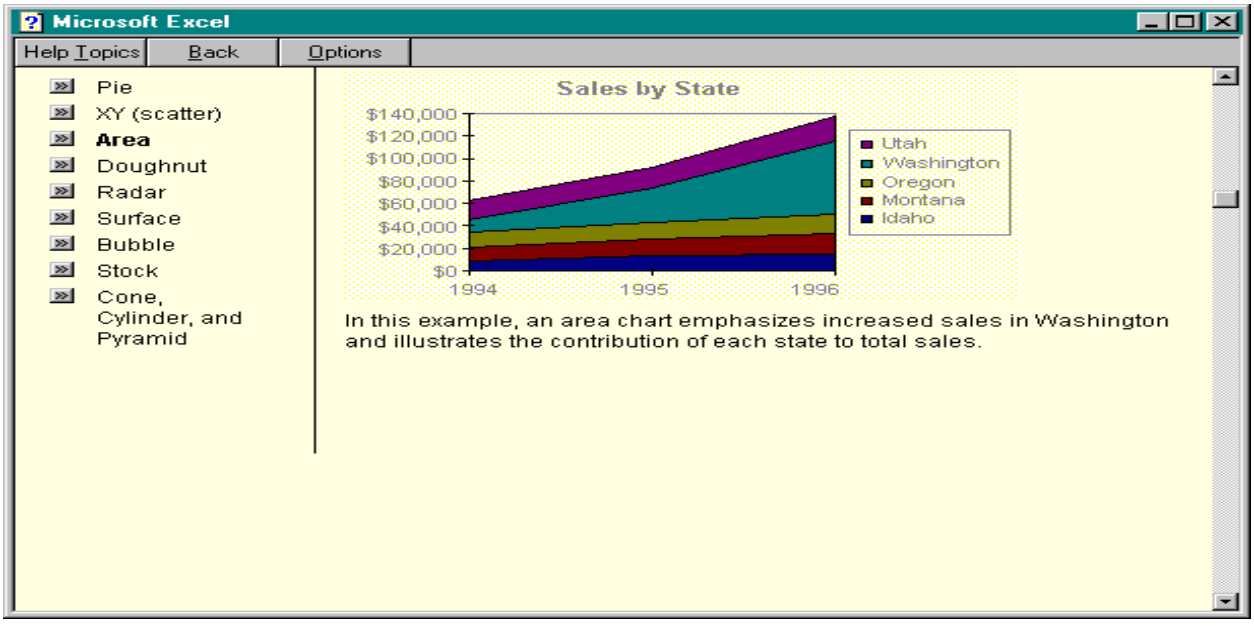
5.6.5 XY (Scatter) Chart

1. A scatter chart is like a line chart but with a greater variety of built-in axes for proportional calibrations.
2. An XY (scatter) chart either shows the relationships among the numeric values in several data series or plots two groups of numbers as one series of xy coordinates.
3. It shows uneven intervals or clusters of data and is commonly used for scientific data.
4. When you arrange your data, place x values in one row or column, and then enter corresponding y values in the adjacent rows or columns.



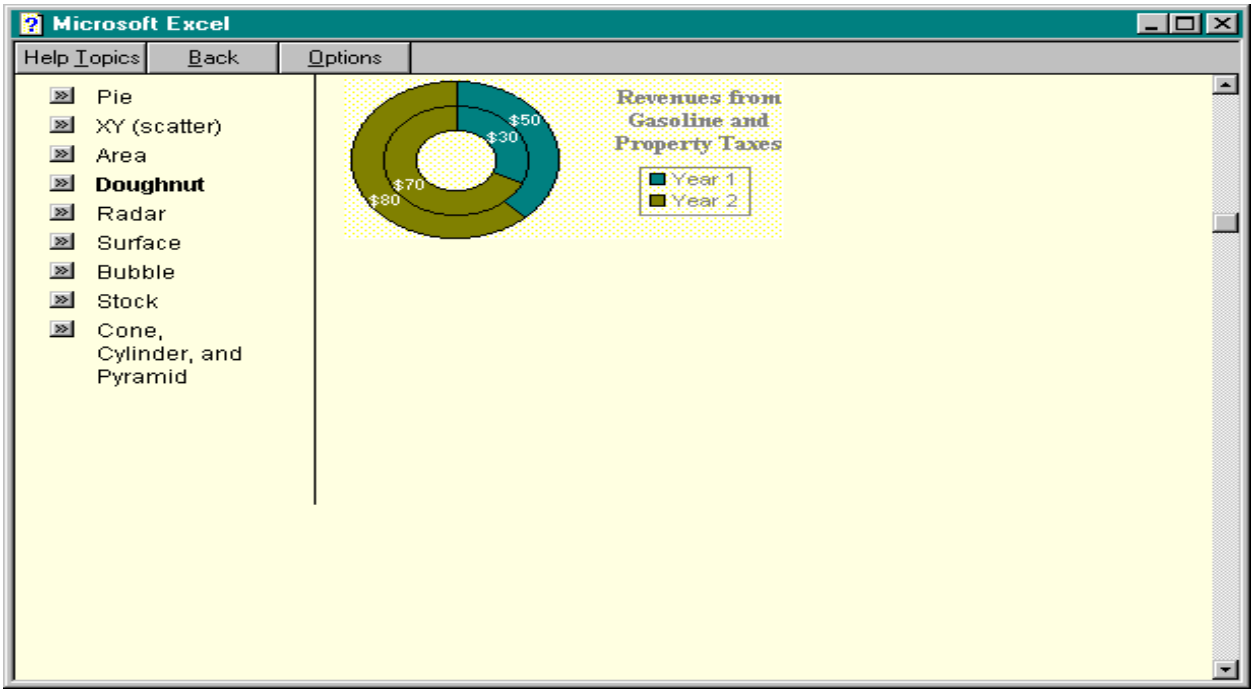
5.6.6 Area Chart

1. An area chart shows the area under a curve.
2. It is usually used to emphasize the magnitude of change over time.
3. By displaying the sum of the plotted values, an area chart also shows the relationship of parts to a whole.
4. For example, an area chart might emphasize increased sales in Washington and illustrate the contribution of each state to total sales



5.6.7 Doughnut Chart

Like a pie chart, a doughnut chart shows the relationship of parts to a whole, but it can contain more than one data series. Each ring of the doughnut chart represents a data series.

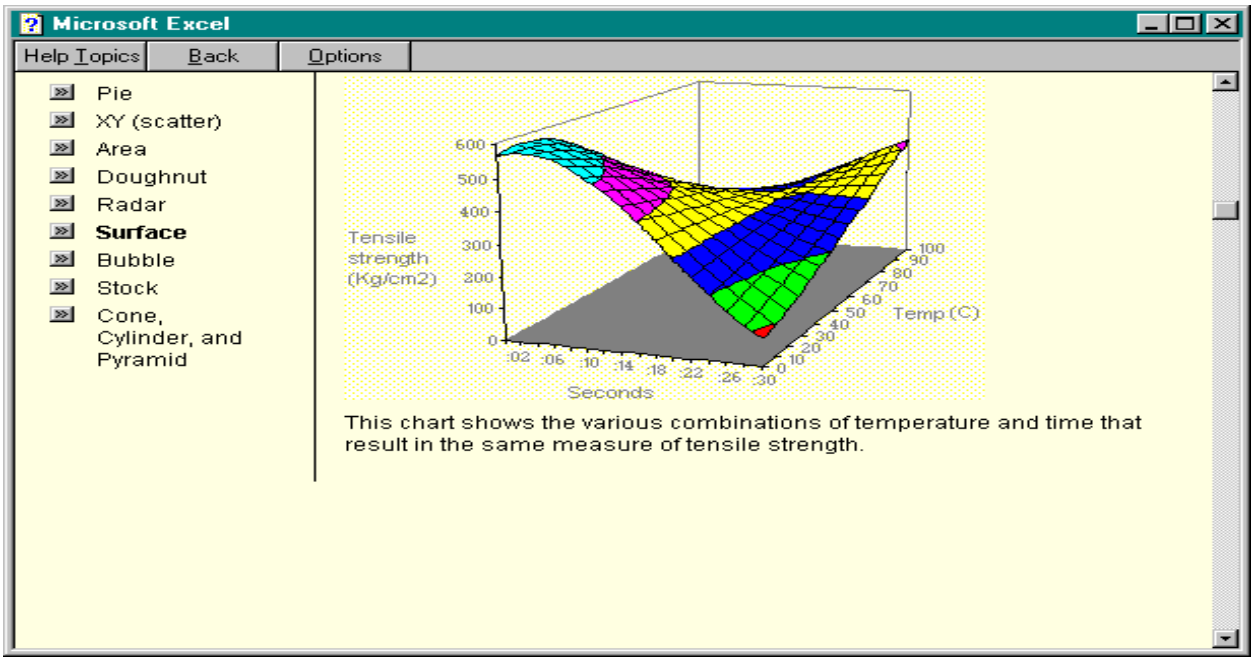


5.6.8 Radar Chart

In a radar chart, each category has its own value axis radiating from the center point. Lines connect all the values in the same series.

5.6.9 Surface Chart

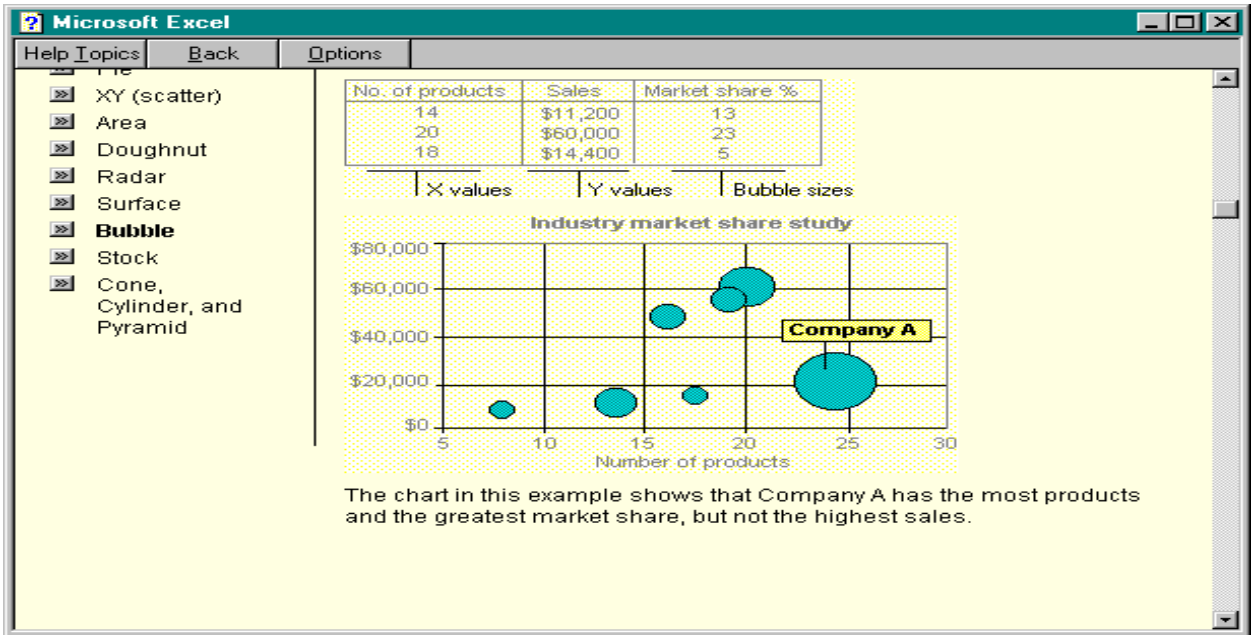
1. A surface chart is useful when you want to find optimum combinations between two sets of data.
2. As in a topographic map, colors and patterns indicate areas that are in the same range of values.



5.6.10 Bubble Chart

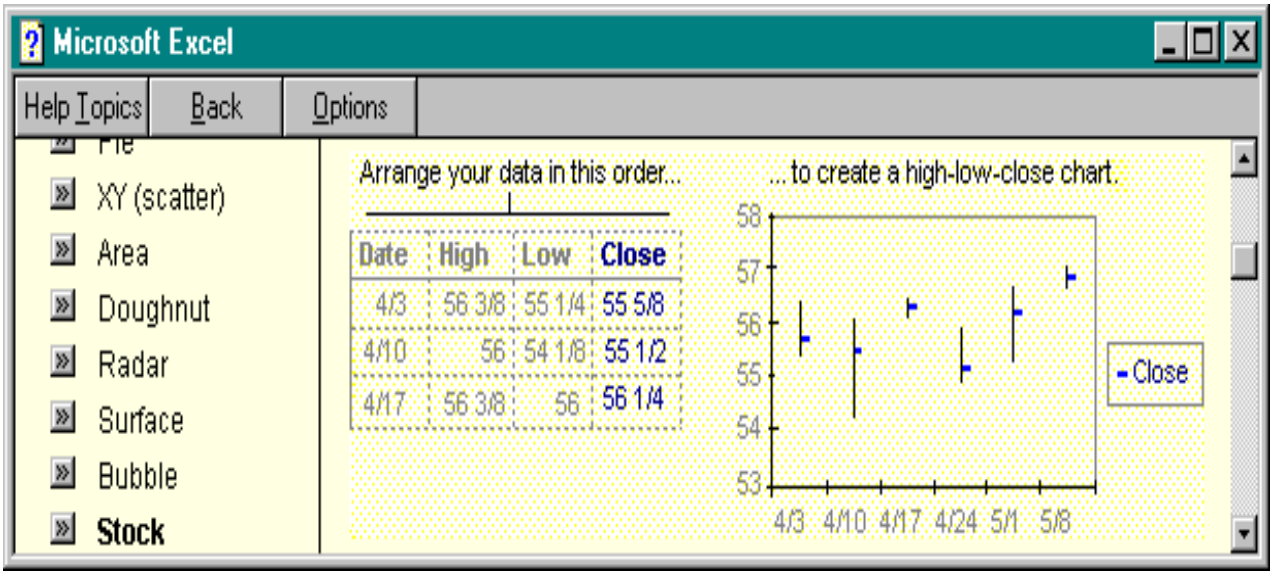
1. A bubble chart is a type of xy (scatter) chart. The size of the data marker indicates the value of a third variable.
2. To arrange your data, place the x values in one row or column, and enter corresponding y values and bubble sizes in the adjacent rows or columns.

a. For example, a chart might show that Company A has the most products and the greatest market share, but not the highest sales.

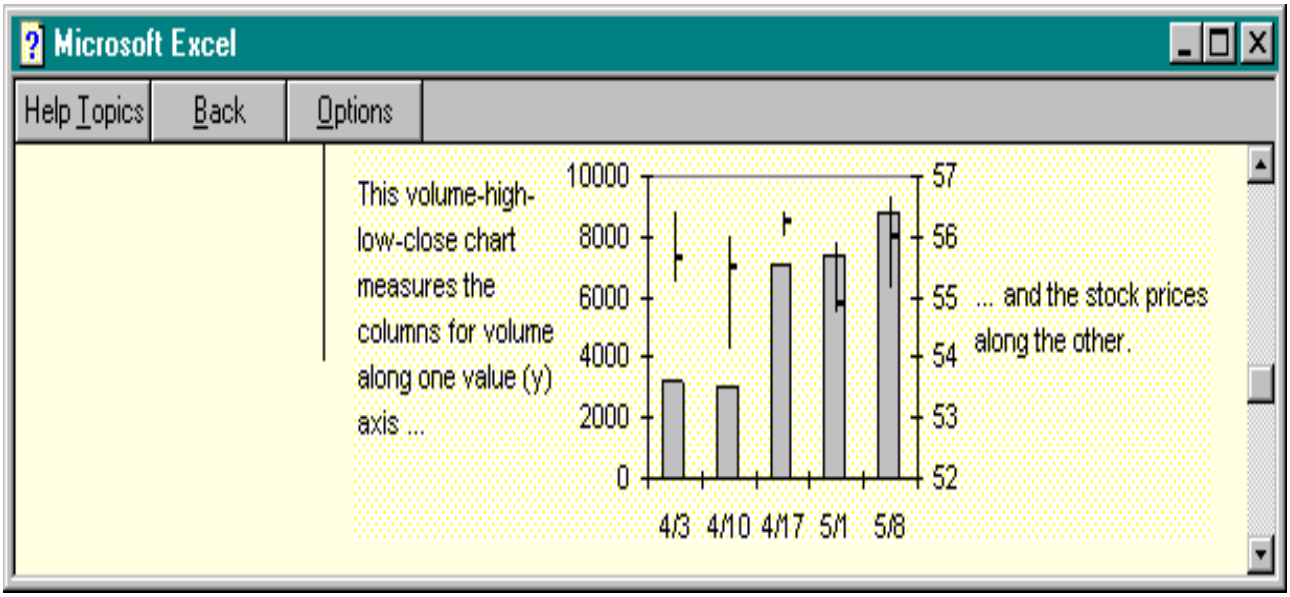


5.6.11 Stock Chart

1. The high-low-close chart is often used to illustrate stock prices.
2. This chart can also be used for scientific data, for example, to indicate temperature changes.
3. You must organize your data in the correct order to create this and other stock charts.

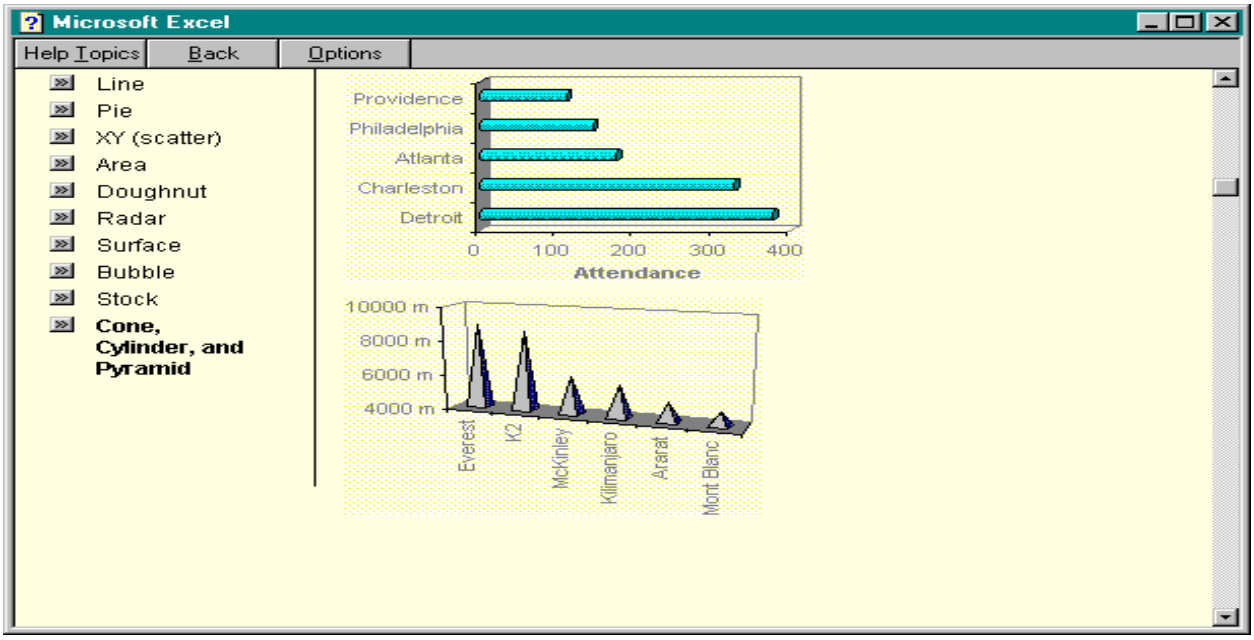


4. A stock chart that measures volume has two value axes: one for the columns that measure volume, the other for the stock prices.
5. You can include volume in a high-low-close or open-high-low-close chart.



5.6.12 Cone, Cylinder, and Pyramid

1. The cone, cylinder, and pyramid data markers are variations of columns and bars. They can add a special effect to 3-D column and bar charts.
2. So instead of having bars and columns you use cones, cylinders, and pyramids.



Self-Check Exercise 5.4

Q1. What are Column Chart.
Q2. What are Bar Charts and Pie Charts? Explain
Q3. What are Doughnut Chart.

5.7 SUMMARY

In this unit, we learnt about the computer graphics and its advantages. We also learned about the different types of graphic presentation. We also learned about how computer is used to draw the bar chart, pie chart, stock chart doughnut chart, cone cylinder pyramid etc.

5.8 GLOSSARY

- **Charts** are used to help people understand numerical data through visualization. Appropriate charts can provide different perspectives, details, overviews, generalizations, and trends in data.
- A **column chart** represents data using vertical bars and is useful for showing changes over time or comparing different items.
- A **bar chart** displays information through horizontal bars, primarily for comparing individual data points.

- A **line chart** plots data points along a graph, either with or without markers and connecting lines, to illustrate trends over time.
- A **pie chart** presents data as a circular graph divided into sectors, showing each category's proportional contribution to the total (100%).
- A **scatter chart** resembles a line chart but offers more flexibility in scaling axes, making it useful for identifying relationships between variables.
- An **area chart** highlights the magnitude of changes over time by shading the space beneath a curve, often emphasizing cumulative values.
- A **doughnut chart** is similar to a pie chart but can represent multiple data series, with each ring corresponding to a different dataset.
- A **radar chart** plots values on separate axes radiating from a central point, connecting them to form a polygonal shape, useful for comparing multiple variables.
- A **surface chart** is ideal for identifying optimal combinations between two datasets, using colors and patterns to represent data ranges, similar to a topographic map.
- A **bubble chart** extends the XY scatter plot by incorporating a third variable, represented by the size of the data points.
- **Cone, cylinder, and pyramid charts** are variations of column and bar charts, offering a three-dimensional effect for visual enhancement.

5.9 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 5.1

Ans. Q1. Refer to Section 5.3

Ans. Q2. Refer to Section 5.3.1

Self-Check Exercise 5.2

Ans. Q1. Refer to Section 5.4

Self-Check Exercise 5.3

Ans. Q1. Refer to Section 5.5

Self-Check Exercise 5.4

Ans. Q1. Refer to Section 5.6.1

Ans. Q2. Refer to Sections 5.6.2 and 5.6.4

Ans. Q3. Refer to Section 5.6.7

5.10 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House.

5.11 TERMINAL QUESTION

Q1. What are computer graphics? What are the advantages of computer graphics?

Q2. Explain the application of computer in presenting bar chart, pie chart, bubble chart and line chart?

Unit-06

CENSUS AND SAMPLE METHOD

STRUCTURE

6.1 INTRODUCTION

When relevant secondary data are unavailable for the specific research problem being investigated, researchers may opt to gather primary data. This involves directly collecting fresh information to address the study's objectives. To achieve this, one can choose between two primary approaches: the **census method** or the **sampling method**. Each of these methods has its own advantages, limitations, and areas of applicability. In this unit, we will explore both methods in detail, examining their significance, differences, and the circumstances under which each is most suitable for data collection.

6.2 LEARNING OBJECTIVES

After going through this unit you will be able to:

- understand the meaning of census and sample methods;
- discuss the merits and demerits of these methods.

6.3 CENSUS METHOD

In the census or complete enumeration survey method, data is gathered from every individual unit within the population, whether it be a person, household, field, shop, or factory, depending on the context. This approach ensures that information is collected from the entire group of interest in a particular study. For instance, to determine the average wage of workers employed in the Indian sugar industry, wage data would be obtained from all workers in the sector. The total wages earned by these workers would then be divided by the number of employees to calculate the average wage. A national population census is another example of this method. Typically, this approach is preferred when the scope of investigation is limited and demands a detailed examination of the population.

6.3.1 Advantages of a Census Survey

The advantages of a census survey are as follow:

- **Intensive Study** – Under census investigation, you must obtain data from each and every unit of the population. Further, it enables the statistician to study more than one aspect of all items of the population. To give you an example, the Indian Government conducts a census investigation once every 10 years. The authorities collect the data regarding the population size, males, and females, education levels, sources of income, religion, etc.
- **Reliable Data** – The data that a statistician collects through a census investigation is more reliable, representative, and accurate. This is because, in a census, the statistician observes every item personally.
- **Suitable Choice** – It is a great choice in situations where the different items of the population are not homogeneous.
- **The basis of various surveys** – Data from a census investigation is used as a basis in various surveys.

6.3.2 Disadvantages of a Census Survey

A census survey also has certain disadvantages. Some of these disadvantages are:

- **Costs** – Since the statistician closely observes each and every item of the population before collecting the data, it makes a census investigation a very costly method of investigation. Usually, government organizations adopt this method to collect detailed data like the population census or agricultural census or the census of industrial protection, etc.
- **Time-consuming** – A census investigation is time-consuming and also requires manpower to collect original data.

- **Workforce-** Owing to the huge volume of data that is collated, a greater number of the workforce (as well as man-hours) is required for completion.
- **Possibilities of Errors** – There are many possibilities of errors in the census investigation method due to non-response, measurement, lack of preciseness of the definition of statistical units or even the personal bias of the investigators.
- **Unsuitability:** Census method is not applicable or suitable if the universe is large. This method is suitable only for a small universe.

Self-Check Exercise 6.1

Q1. What is meant by Census Survey.

Q2. Write the merits and demerits of Census Survey.

6.4 SAMPLING

Sampling is a methodological approach used to gain insights into a larger population by analyzing a representative subset of its units. Instead of conducting a comprehensive study of every individual element within a given population, researchers select a sample that serves as a basis for making inferences about the entire group. A sample is essentially a smaller, manageable portion of the population that is examined to derive meaningful conclusions.

The process of sampling consists of three key components:

a) **Selection of the Sample** – Identifying and choosing a subset of the population that accurately represents the whole.

b) **Data Collection** – Gathering relevant information from the selected sample through various techniques such as surveys, observations, or experiments.

c) **Inference and Analysis** – Drawing conclusions about the broader population based on the collected data.

These three elements are interconnected and cannot be treated as independent stages, as each step influences the other. The accuracy and reliability of the results depend on the selection process, data collection methods, and the statistical techniques used for analysis. Sampling does not imply arbitrary selection; rather, it follows specific principles and methodologies to ensure that the sample is representative of the entire population.

Although significant advancements in sampling theory have been made in recent years, the fundamental concept of sampling is ancient. Humans have long used sampling techniques in everyday life. For instance, farmers have traditionally assessed the quality of a grain harvest by examining just a handful of grains. Similarly, a homemaker checks the doneness of a pot of rice by testing only a few grains. Medical professionals draw conclusions about a patient's overall health by analyzing a small sample of blood. Businesspersons evaluate the quality of materials before placing large orders by inspecting only a small portion of the stock. Educators assess whether a class understands a lesson by questioning a few students.

Sampling is widely used across various domains, both consciously and unconsciously, as a practical means of deriving conclusions without the need for exhaustive data collection. It is important to recognize that a sample is not examined in isolation; rather, it is studied to infer characteristics about the entire population. The values obtained from sample analysis, such as averages or measures of dispersion, are referred to as 'statistics,' whereas the corresponding values for the entire population are termed 'parameters.'

In essence, sampling serves as a crucial tool in research and decision-making processes, enabling efficient data analysis while minimizing costs and effort. By studying a well-chosen sample, one can make informed judgments about a larger population with a reasonable degree of accuracy.

6.4.1 Theoretical Basis of Sampling

Through the analysis of sample data, we can make predictions and generalize the behavior of large-scale phenomena. This is feasible because no statistical population consists of elements that vary without limitation. For instance, wheat exhibits some variation in characteristics such as color, protein content, length, and weight, yet it remains identifiable as wheat. Likewise, apples from the same tree may differ in size, color, taste, and weight, but they are still recognizable as apples. This demonstrates that while variability is an inherent feature of mass data, every population possesses distinct characteristics with a limited range of variation. As a result, a relatively small, unbiased random sample can effectively represent the attributes of the entire population.

There are two important laws on which the theory of sampling is based:

6.4.1.1 Law of Statistical Regularity

6.4.1.2 Law of Inertia of Large Numbers

6.4.1.1 Law of Statistical Regularity

This principle is rooted in probability theory and states that when a sufficiently large number of items are randomly selected from a larger population, they are likely to reflect the characteristics of that population. As King explains, "The law of statistical regularity asserts that a reasonably large sample, chosen at random from a larger group, will, on average, exhibit the same properties as the group itself."

In essence, this principle emphasizes that a randomly drawn sample is expected to closely resemble the overall population. This highlights the importance of selecting samples randomly to ensure unbiased representation. A truly random selection means that every item in the population has an equal chance of being included in the sample, eliminating any subjective bias. When this condition is met, studying a subset of the population can provide reliable insights into the entire group.

The practical significance of this principle lies in its ability to simplify research efforts by reducing the workload required to draw conclusions about a large population. For instance, to estimate the average height of students at Himachal Pradesh University, it is unnecessary to measure every student's height. Instead, a random

selection of students from different colleges can be measured, and their average height can be used to approximate the overall university average.

It is important to acknowledge that sample-based results may not be identical to the actual population values. Since a sample represents only a portion of the entire population, minor variations are inevitable. For example, while a census may determine the average height of Delhi University students to be 160 cm, a sample study may yield 159 cm or 161 cm. Achieving an exact match would be coincidental, but if the sample is well-chosen, the difference in results will be minimal.

6.4.1.2 Law of Inertia of Large Numbers

The Law of Inertia of Large Numbers is a direct outcome of the Law of Statistical Regularity and plays a crucial role in sampling theory. It suggests that, under similar conditions, increasing the sample size leads to more accurate results. This is because larger numbers tend to exhibit greater stability compared to smaller ones. When a sample is large, variations in individual components tend to balance each other out, making fluctuations in the overall outcome minimal.

For instance, if a coin is flipped 10 times, the expected outcome is five heads and five tails. However, due to the limited number of trials, the actual result might differ—such as obtaining 9 heads and 1 tail or 7 heads and 3 tails. On the other hand, if the coin is tossed 1,000 times, the likelihood of achieving close to 500 heads and 500 tails is significantly higher. This occurs because the greater number of trials increases the probability of compensatory variations, ensuring that any temporary bias in one direction is neutralized over multiple repetitions.

Similarly, when analyzing rice production trends over several years, using data from just one or two states might show significant fluctuations due to localized factors. However, if production data from all Indian states are considered, the overall variation is likely to be smaller. This does not imply that production levels remain unchanged each year but rather that the fluctuations across different states tend to offset each other, resulting in a more stable aggregate trend.

6.4.2 Essentials of Sampling

If the sample results are to have nay worthwhile meaning, it is necessary that a sample possesses the following essentials:

(i) **Representativeness**: A sample should be so selected that it truly represents the universe otherwise the results obtained may be misleading. To ensure representativeness the random method of selection should be used.

(ii) **Adequacy:** The size of sample should be adequate; otherwise it may not represent the characteristics of the universe.

(iii) **Independence:** All items of the sample should be selected independently of one another and all items of the universe should have the same chance of being selected in the sample. By independence of selection we mean that the selection

of a particular item in one draw has influence on the probabilities of selection in any other draw.

(iv) **Homogeneity**: When we talk of homogeneity we mean that there is no basic difference in the nature of units of the universe and that of the sample. If two samples from the same universe are taken they should give more or less the same unit.

(v) **No bias and prejudices:** The selection of the sample should be objective. Sample should be free from bias and prejudices. Then only dependable result can be achieved. Investigator has to be very cautious in this task.

(vi) **Conformity with the subject-matter and meant:** In sample methods the representative units selected, should be as per the subject matter and means.

(vii) **Accuracy:** Accuracy is defined as the degree to which bias is absent from the sample. An accurate sample is the one which exactly represents the population.

(viii) **Precision:** The sample must yield precise estimate. Precision is measured by standard error.

6.4.3 Advantages of Sampling

i) **Time Efficiency** – Since only a subset of items is collected and analyzed, the sampling method significantly reduces the time required to obtain results. This is particularly useful when quick decision-making is necessary.

ii) **Cost Reduction** – Sampling minimizes expenses as it involves studying a limited number of selected items. This leads to lower financial costs and reduced labor hours, making it especially beneficial for resource-constrained economies.

iii) **Enhanced Reliability** – The accuracy of results tends to be higher because:

a) The likelihood of statistical errors is reduced, and any existing errors can be identified and controlled.

b) Skilled professionals can focus on analyzing a smaller dataset, applying advanced techniques to ensure more precise outcomes.

iv) **Detailed Analysis** – Since sampling conserves time, money, and effort, it allows for the collection of more in-depth and comprehensive data.

v) **Practical Feasibility** – In certain cases, sampling is the only viable option. For instance, testing the strength of bricks in a factory through a complete census would render all bricks unusable, making a full-scale assessment impractical. Additionally, when dealing with an infinite population, sampling remains the sole feasible method.

vi) **Simplified Administration** – Organizing and managing a sample survey is easier compared to conducting a full census, making implementation more efficient.

vii) **Scientific Validity** – Sampling follows a structured, scientific approach, ensuring the justification of resources spent on data collection and analysis.

viii) **Higher Accuracy** – Compared to the census method, sampling often yields more precise results due to better resource allocation and targeted analysis.

It is very important to note that the aim of sampling studies is to obtain maximum information about the phenomena under study with least sacrifice of money, time and energy. The purpose of sampling is to get information about the population from the sample. For example, a doctor examines few drops of blood and draws conclusions about the whole blood. We can obtain a large variety of information about the phenomena to which the sample relates. And, this helps us to have an idea about similar information relating to the universe. For example, when we go to the market we examine a sample of rice (a handful of rice) from the lot, form an idea about the quality and decide whether the quality is acceptable or not. Another example, we meet a person for a while, talk with him and form opinion about his character. In all these examples, we adopt sampling technique. Our knowledge, our attitude and our actions are based to a large extent on samples. The theory which helps us is studying samples is known as theory of sampling. Logic theory of sampling is the logic of induction i.e. from the study of a sample, one tries to infer about the population. Thus, the aim of sampling studies is to obtain the best possible values of the parameters. The population measures for example, mean, standard deviation etc. called parameters, while the measures obtained from sample are called statistics.

6.4.4 Disadvantages of Sampling

i) **Possibility of Misleading Conclusions**: If a sampling process is not meticulously designed and implemented, the findings may be inaccurate and lead to incorrect conclusions.

ii) **Lack of Representativeness**: A sample must accurately reflect the entire population for the results to be applicable. If the sample is not representative, the conclusions drawn may be unreliable and misleading.

iii) **Shortage of Skilled Professionals**: The effectiveness of a sample survey depends on experts who can plan, conduct, and analyze the data. A lack of expertise may compromise the accuracy and reliability of the results.

iv) **Complexity and Cost**: In some cases, a well-structured sampling plan may require more time, labor, and financial resources than a complete enumeration (census), making it a less practical choice.

v) **Organizational Challenges**: Conducting a sample-based study involves several logistical and administrative difficulties, which can affect its efficiency.

vi) **Subjectivity and Bias**: Personal biases or preferences in selecting sampling techniques and sample units may influence the findings, reducing the objectivity of the results.

vii) **Inappropriate Sample Size**: If the sample size is too small or too large, it may not accurately reflect the characteristics of the population, leading to flawed conclusions.

viii) **Limitations in Full Coverage**: When information is required for every individual or unit within a population, a complete enumeration survey is preferable to sampling.

Despite these limitations, sampling remains a valuable research tool when conducted using a scientific approach. As Frederick F. Stephen aptly states, “Samples are like medicines. They can be harmful when used carelessly or without adequate understanding. However, when applied with caution and knowledge, they yield reliable results.” Similarly, Professor Chou emphasizes that sampling is a method of studying a population by analyzing a representative subset, allowing researchers to make generalizations based on the findings.

Self-Check Exercise 6.2

Q1. What do you mean by sampling?

Q2. Discuss the characteristics of a good sample.

Q3. Write the merits and demerits of sampling.

6.5 DIFFERENCE BETWEEN CENSUS AND SAMPLE SURVEY

The main differences between census and sample survey are as below:

Census Method (Survey)	Sample Method
In this survey, information is collected from each and every unit of the population.	In this survey, information is collected from a few selected units of the population.
The Census Method is very expensive and time-consuming.	The sample Method is less expensive and less time-consuming.
The Census Method is suitable where the field of investigation is small.	The sample method is suitable where the field of investigation is large.
The Census Method is more accurate and reliable.	The Sample Method is less accurate and less reliable.
The Census Method rules out the possibility of any personal biases.	Sample Method holds the chance of personal biases in the selection of samples.

Self-Check Exercise 6.3

Q1. Distinguish between census and sampling methods

6.6 SUMMARY

Census and sample surveys essentially relate to the statistical collection of data across various areas and sectors pertaining to the particular subject matter or inquiry. In Census Method, each and every item in the universe is selected for the data collection. The selected data might constitute a particular place, a group of people, or any specific locality that is the complete set of items and which are of interest in any particular situation. The sample method consisting of the selecting for study, a portion of the Universe with a view to draw conclusions about the Universe or population is known as sampling. These two methods has their own merits and demerits. On the basis of nature of study, availability of funds, time, and manpower, we may choose one of these two methods to collect the information for our study.

6.7 GLOSSARY

- **Census Survey:** In a census or complete enumeration survey, data is gathered from every individual unit within a population, including persons, households, fields, shops, factories, and other relevant entities.
- **Sampling:** Sampling involves selecting a subset of the total population for study to make inferences about the entire group. This method helps in drawing conclusions about the overall population based on a representative portion.
- **Law of Statistical Regularity:** This principle states that when a sufficiently large number of items are randomly selected from a larger population, they are likely to exhibit the same characteristics as the entire group on average.
- **Law of Inertia of Large Numbers:** This law suggests that, under similar conditions, the accuracy of results improves as the sample size increases. Larger samples tend to be more stable and reliable compared to smaller ones.

6.8 ANSWERS TO SELF-CHECK EXERCISES

Self-Check Exercise 6.1

Ans. Q1. Refer to Section 6.3

Ans. Q2. Refer to Sections 6.3.1 and 6.3.2

Self-Check Exercise 6.2

Ans. Q1. Refer to Section 6.4

Ans. Q2. Refer to Section 6.4.3

Ans. Q3. Refer to Sections 6.4.4 and 6.4.5

Self-Check Exercise 6.3

Ans. Q1. Refer to Section 6.5

6.9 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of

Statistics, Kitab Mahal, New Delhi.

- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House.

6.10 TERMINAL QUESTIONS

Q1. What is census survey. Write its merits and demerits.

Q2. What is the significance of sampling in research?

Q3. What is sampling. Write its merits and demerits.

Q4. What are meant by census and sampling methods. Distinguish between census and sampling methods.

Unit-07

SAMPLING METHODS

STRUCTURE

7.1 INTRODUCTION

In the last unit, we have studied about the meaning of sample methods, its merits and demerits. In this unit, we will discuss the various types of sampling methods. The random and non-random sampling methods and their merits and demerits all have been discussed in detail in this unit.

7.2 LEARNING OBJECTIVES

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

- understand the meaning of a sample
- study the various sampling methods
- discuss the different types of probability or random sampling
- explain the various types of non-probability or non-random sampling

7.3 TYPES OF SAMPLING

In survey research, sampling refers to selecting a subset of a population to represent the entire group. This approach makes large-scale studies more feasible in terms of cost and time, as it involves analyzing a smaller, yet representative, portion of the population. Essentially, sampling enables researchers to draw conclusions about a population based on a selected sample rather than studying every individual. Various sampling methods exist, and the choice of technique depends on the research objective. These methods are broadly categorized into two groups:

7.3.1 Random Sampling Method (Probability Sampling)

7.3.1.1 Simple or Unrestricted Random Sampling

7.3.1.2 Restricted Random Sampling

(i) Stratified Sampling

(ii) Systematic Sampling

(iii) Cluster Sampling

7.3.2 Non-Random Sampling (Non-probability Sampling)

7.3.2.1 Judgment or Purposive Sampling

7.3.2.2 Quota Sampling

7.3.2.3 Convenience Sampling

7.3.2.4 Snow-ball Sampling

7.3.1 Random Sampling (Probability Sampling)

Random sampling refers to a method where every item in the population has an equal and known chance of being chosen. Dr. Yates defines it as a process where "every member of a parent population has had equal chances of being included." Similarly, Harper describes it as a sample selection approach ensuring that every element in the population has an equal opportunity of being part of the sample.

7.3.1.1 Simple or Unrestricted Random Sampling

This technique ensures that each unit in the population has an equal and independent probability of being included in the sample. Various methods are employed to select a sample randomly, including:

(i) Lottery Method: This is one of the simplest and most widely used techniques. Each item in the population is assigned a number and written on individual slips of identical

size, shape, and color. These slips are then shuffled in a container, and a blindfolded selection is made. The required number of slips is drawn randomly to form the sample.

For instance, if we need to select five students from a group of fifty, we would write all fifty names on identical slips, mix them, and then randomly pick five slips. This approach is also known as unrestricted random sampling since there are no restrictions on the selection process. It is commonly applied in lottery draws. However, if the population is infinite, this method is not feasible. Additionally, if the slips vary in size or shape, the selection process may be biased.

(ii) Table of Random Numbers: When dealing with an infinite population, the lottery method becomes impractical. In such cases, a table of random numbers serves as an alternative. Several well-known tables of random numbers exist, with one of the earliest developed by Prof. L.H.C. Tippett in 1927, derived from the British Census Report. His table consists of 10,400 four-digit numbers, totaling 41,600 digits. Other notable random number tables include those by Fisher and Yates (1938), which contain 15,000 digits grouped in pairs, Kendall and B.B. Smith (1939) with 100,000 digits organized into 25,000 sets of four-digit numbers, and the Rand Corporation (1955), which features 200,000 five-digit random numbers.

Advantages of Random Sampling

i) **Scientific Approach:** The method minimizes personal bias.

ii) **Greater Representativeness:** As sample size increases, it becomes more representative due to the Law of Large Numbers and the Law of Statistical Regularity.

iii) **Measurable Sampling Error:** Errors in sampling can be estimated and analyzed.

iv) **Applicability of Probability Theory:** Probability principles apply when a sample is selected randomly.

v) **Cost-Effective:** It saves time, effort, and financial resources.

Limitations of Random Sampling

i) **Requirement of a Complete Population List:** A comprehensive and updated list of the population is necessary, but such lists may not always be available.

ii) **Lack of Representativeness in Small Samples:** If the sample size is too small, it may not accurately represent the entire population.

iii) **Inapplicability for Large Distributions:** When the population is highly dispersed, using this method becomes challenging.

7.3.1.2 Restricted Random Sampling

(i) Stratified sampling: Stratified sampling is used when a population consists of diverse segments or strata with respect to a particular characteristic under study. In this method, the population is initially divided into several homogeneous sub-groups, known as strata, and a random sample is then selected from each stratum. Stratified random sampling is classified into two types: proportional and non-proportional sampling. In proportional stratified sampling, the selection of samples from each subgroup is proportionate to their size in the overall population. Larger subgroups contribute more to the sample, while smaller ones contribute less. In contrast, non-proportional sampling gives equal representation to all strata, regardless of their actual size in the population.

Advantages:

i) This method enhances the representativeness of the sample.

ii) It provides higher accuracy in results.

iii) Since the population is categorized into sub-groups, managing the sampling process becomes easier.

iv) It reduces time and costs, particularly when there is a geographical concentration.

v) It is particularly useful when the original population distribution is highly skewed.

vi) For heterogeneous populations, it can produce more reliable outcomes.

Disadvantages:

i) The process of dividing a population into homogeneous strata requires significant time, resources, and expertise.

ii) Inaccurate stratification may introduce bias, and overlapping strata may result in an unrepresentative sample.

(ii) Systematic sampling: Also referred to as quasi-random sampling, systematic sampling is employed when a comprehensive list of the population is available. Under this method, the items are arranged in numerical, alphabetical, geographical, or any other logical order. A sample is then selected by picking every Kth item from the sampling frame, where K is the sampling interval. For instance, if a sample of 10 students is required from a population of 100 students, the sampling interval (K) is calculated as follows:

K = Sampling interval

N = Size of universe

n = Sample size

in the above example k =10. 10 is the sampling interval. Every 10th student will be taken as sample, i.e., 10th 20th, 30th, and so on.

Advantages:

i) This method is straightforward and easy to implement.

ii) It significantly reduces time and effort.

iii) If executed carefully, it can yield satisfactory results and is applicable even for large populations.

Disadvantages:

i) It may not always provide a fully representative sample.

ii) There is a risk of personal bias influencing the selection process.

(iii) Cluster sampling or multistage sampling. Cluster sampling, also known as multistage sampling, involves selecting samples through multiple stages. The entire population is initially divided into primary sampling units, which are then further subdivided into smaller units. This process continues until a manageable sample size is achieved. For example, if we need to select 5,000 students from Madhya Pradesh, we could follow these steps:

- **First stage:** Select universities within Madhya Pradesh.
- **Second stage:** Choose a sample of colleges from the selected universities.
- **Third stage:** Select students from these colleges.

Advantages:

i) This method provides flexibility in sampling.

ii) It is particularly useful for large-scale surveys where compiling a complete list of population units is impractical, expensive, or time-consuming.

iii) In developing regions with limited or incomplete data records, this method can be highly beneficial.

Disadvantages:

i) It is generally less accurate compared to other sampling techniques.

7.3.2 Non-random Sampling Method (Non-Probability Sampling)

7.3.2.1 Judgment sampling (Purposive or Deliberate): In judgment sampling, also known as purposive or deliberate sampling, the investigator exercises discretion in selecting or rejecting specific items for the study. The choice of sample elements is based on the researcher's judgment, making their role crucial in data collection. For instance, if an investigator needs to select five students from a class of fifty B.Com. students to analyze movie-watching habits, they would choose the individuals they believe best represent the class.

Advantages

i) It is a straightforward and easy-to-use method.

ii) It helps in obtaining a more representative sample.

iii) This approach is particularly useful in policy-making and decision-making, often employed by executives and public officials to address urgent issues.

Disadvantages

i) The investigator's personal bias may lead to a non-representative sample.

ii) Identifying accurate sampling errors is challenging.

iii) The estimates obtained may lack precision.

iv) Results from this sampling method are not easily comparable with other sampling studies.

7.3.2.2 Quota Sampling

Quota sampling is a non-random sampling technique similar to stratified sampling. It is commonly used in the United States for public opinion surveys and consumer research. In this method, the population is divided into subgroups (quotas) based on certain characteristics. Each data collector is assigned a specific quota of individuals to interview. The sample selection relies on personal judgment, making it a combination of stratified and purposive sampling, which allows researchers to leverage the benefits of both methods. This method is time- and cost-efficient. If skilled investigators conduct the sampling, it can yield reliable results. However, personal bias may influence the selection process. Since it is not a random sampling method, estimating sampling errors is not possible.

7.3.2.3 Convenience Sampling (Chunk Sampling)

Convenience sampling, also known as chunk sampling, involves selecting population units based on ease of access. The sample is a convenient subset of the population rather than being randomly chosen.

Advantages

i) It is useful when the total population is not clearly defined.

ii) It does not require a well-defined sampling unit.

iii) It is applicable when a comprehensive population list is unavailable.

For example, samples obtained from automobile registration records, telephone directories, or social media platforms fall under this category. However, convenience sampling is prone to bias, leading to non-representative and unreliable results. Despite these limitations, it is often used for preliminary or exploratory research.

7.3.2.4 Snow-ball Sampling:

Snowball sampling, also called chain referral sampling, is a non-probability sampling technique frequently used in sociological and statistical research. In this method, existing participants help recruit new participants from their social networks. As the sample expands, more data becomes available, much like a snowball increasing in size as it rolls downhill.

This technique is particularly useful for studying hard-to-reach or hidden populations, such as drug users, sex workers, or undocumented workers. Since the sampling process is not based on a predefined frame, results may be biased. People with larger social circles are more likely to be included in the sample.

Snowball sampling relies on initial informants to identify and refer other potential participants who meet the study's criteria. While this approach is valuable for building networks and expanding participation, its effectiveness heavily depends on the credibility and influence of the initial contacts.

Advantages

i) **Access to hidden populations** – This method allows researchers to reach groups that might otherwise be difficult to include in a study.

ii) **Targeting specific demographics** – It enables researchers to find individuals within niche populations that lack an accessible list or database.

Disadvantages

i) **Community bias** – The sample's composition depends heavily on the first participants, which may introduce significant bias.

ii) **Lack of randomness** – Since it does not follow a random selection process, the results may not be truly representative.

iii) **Uncertain sample size** – There is no clear way to determine the total size of the target population.

iv) **Potential misrepresentation** – The sample may not accurately reflect the broader population, as the recruitment process is influenced by social connections.

Despite these challenges, snowball sampling remains a valuable method in social research, particularly for studying populations that are difficult to access through traditional sampling techniques.

Self-Check Exercise 7.1

Q1. What is meant by random Sampling

Q2. What is Quota Sampling

Q3. What is meant by Snow-ball Sampling

7.4 CRITERIA FOR CHOICE OF SAMPLING METHODS

The reliability of research outcomes and the accuracy of findings largely depend on the suitability of the chosen sampling design. Various constraints and practical challenges necessitate careful consideration of the following factors:

i) **Purpose of the Study**: The selection of a sampling method is influenced by the research objective. If the goal is to generalize findings, probability sampling techniques should be used. However, if the study aims to explore the nature of a problem, non-probability sampling methods are more appropriate.

ii) **Statistical Measurability**: When research requires statistical inferences, it is essential to use sampling methods such as simple random sampling or stratified random sampling. Since statistical inference involves estimating sampling errors, only probability sampling allows for such calculations.

iii) **Required Precision Level**: The choice of a sampling method also depends on the level of precision needed. When a high degree of accuracy is required, probability sampling should be applied. If a lower level of precision is acceptable, non-random sampling methods may suffice.

iv) **Availability of Population Data**: If comprehensive data about the population exist, it is preferable to use a probability sampling method. Conversely, when no population list or prior information is available, non-probability sampling is a more practical choice.

v) **Population Characteristics**: In cases where the population is homogeneous, even simple random sampling can yield a representative sample. However, for heterogeneous populations, stratified random sampling is a more effective approach.

vi) **Geographical Scope of the Study**: If the study covers a large geographic area with a substantial population, multi-stage or cluster sampling is appropriate. On the other hand, for smaller populations and regions, single-stage probability sampling methods can be utilized.

vii) **Financial Constraints**: The available budget plays a crucial role in determining the sampling method. When financial resources are limited, methods like quota or judgment sampling are more feasible. If funding is not a constraint, the researcher can opt for the most suitable sampling technique based on study objectives.

viii) **Time Constraints**: The timeframe within which the research must be completed influences the choice of sampling methods. When time is limited, less time-intensive approaches like simple random sampling are preferable over more complex methods such as stratified sampling.

Self-Check Exercise 7.2

Q1. What are the criteria for the choice of sampling methods?

7.5 The factors which determine the sample size

Sample size here means how many units of the population should be selected for the study? There is no clear cut idea about the required size of sample. As a general rule, the simple must be of an optimum size. It should not very large and too small. Sample size should be large enough to give a precision of the results. Hence size of the sample should be determined by a researcher keeping the following points.

i) **Nature of the population:** Population may be homogenous or heterogeneous. If the population contains a lot of heterogeneous population then large sample size of is required. Small size of population is sufficient if there is homogeneity in thoughts.

ii) **Size of the population:** Depending upon the size of population, the size of sample has to be selected. If the population is very small, the sample size would be small and vice versa.

iii) **Resources available:** The size of sample depends upon the amount of resources available for the study. With sufficient time and large volume of funds available, the sample size could be large, otherwise should be small.

iv) **The extent of accuracy desired:** If the standard of accuracy is to be kept high, we need a large sample and vice versa.

v) **Types of sampling:** Sampling methods is an important part in determining the size of the sample. Depending upon the method sampling used, the size of sample will be decided.

vi) **Nature of study:** If items are to be intensively and continuously studied, the sample should be small. In case of technical studies, small size of sample is appropriate. In case of extensive and one time studies, the size of sample should be large.

vii) **Nature of respondents:** The nature of respondents will influence the sample size. If the respondents are literate the size of sample could be smaller. If the respondents are illiterate, the size of sample should be large.

Self-Check Exercise 7.3

Q1. What determine the sample size?

7.6 Sampling and Non-Sampling Errors

7.6 Sampling and Non-Sampling Errors

A sample survey involves studying a small fraction of the entire population and drawing conclusions based on that subset. Naturally, this approach is prone to certain inaccuracies, commonly referred to as sampling errors or sampling fluctuations. In a complete census, these errors would typically be eliminated.

7.6.1 Sampling Errors

Sampling errors arise when data is collected using a sample rather than the entire population. Even in a random sampling process, the selected sample may not always perfectly represent the population. This is because no sample is a flawless miniature of the entire population. However, these errors can be controlled to a certain extent.

Sampling errors can be categorized into two types:

i. Biased Errors: Biased errors occur due to personal biases or preferences in selecting a sampling method. For example, opting for purposive sampling instead of simple random sampling can introduce bias. Such errors, also referred to as cumulative or non-compensating errors, tend to persist regardless of sample size and may even increase as the sample grows.

These errors may arise due to:

i) **Flawed Selection Process:** Bias can enter through improper sampling techniques, such as purposive selection, random selection done carelessly, substituting sampled items, or incomplete data collection.

ii) **Errors in Data Collection:** Inaccuracies may occur during data gathering due to poor problem formulation, incorrect population definitions, inappropriate decision-making, lack of a proper sampling frame, poorly designed questionnaires, untrained interviewers, respondent memory failure, disorganized data collection, or errors in data editing and coding.

iii) **Incorrect Analytical Methods:** Bias can also result from improper data analysis. Employing appropriate analytical techniques helps in mitigating such errors.

To minimize biases and enhance sampling design, the following measures can be taken:

- Clearly define and manage the research problem
- Conduct detailed studies to identify and report methodological biases
- Systematically document related research
- Allocate sufficient resources for data collection
- Perform thorough pre-testing
- Integrate complementary research methods
- Conduct replication studies

ii. Unbiased Errors: Unbiased errors arise due to random variations between the sample and the population. These are also known as random sampling errors. Unlike biased errors, random errors decrease as the sample size increases. Since these errors tend to balance out, they are often referred to as non-cumulative or compensating errors.

7.6.2 Non-Sampling Errors

Non-sampling errors can occur in any survey, whether it involves a complete population count or a sample-based approach. These errors include biases and mistakes that are independent of the sampling process. Some key factors contributing to non-sampling errors are:

- Unclear population definition
- Poorly structured questionnaires
- Ambiguity in the information being sought
- Inappropriate statistical units
- Inaccurate or unsuitable interviewing techniques
- Errors in observation or measurement
- Mistakes in data processing, including coding, entry, verification, and tabulation
- Errors in result presentation and publication

Unlike sampling errors, non-sampling errors tend to increase with sample size. Therefore, efforts should be made to control and minimize them to maintain data accuracy and reliability.

Self-Check Exercise 7.4

Q1. What is meant by Sampling Errors

Q2. What is Non-sampling Errors

7.7 SUMMARY

Sampling is the process of selecting a subset from a larger population, known as the universe, to draw conclusions about the entire group. Sampling methods can be broadly categorized into Probability (Random) Sampling and Non-Probability (Non-Random) Sampling. Probability sampling involves selecting samples using a random process, ensuring that every element in the population has an equal and independent chance of being chosen. The simplest form of this method is Simple Random Sampling, where each unit is selected purely by chance. Stratified Random Sampling, also known as proportional or quota sampling, involves dividing the population into homogeneous subgroups and then selecting a random sample from each subgroup. This method is particularly useful for large and diverse populations.

When the population is spread over a vast geographical area and a comprehensive list of all elements is unavailable, Cluster Sampling is often employed due to its convenience and efficiency. Multistage Sampling is another technique where sampling occurs in multiple stages, with each stage involving a cluster of elements from the next stage. Different random sampling techniques are applied at each stage, making it ideal for geographically dispersed populations without a well-defined sampling frame. Within multistage sampling, subsequent sampling processes are referred to as sub-sampling.

In some cases, random sampling may not be feasible, practical, or suitable for research objectives. In such situations, Non-Probability Sampling is used. This method is divided into two main types: Accidental Sampling, where subjects are chosen based on availability, and Purposive Sampling, where participants are selected based on specific criteria. Sampling errors arise due to the sampling process itself and can be classified as biased or unbiased errors. Additionally, Non-Sampling Errors can occur in any survey, whether based on complete enumeration or sampling. These errors may include biases, inaccuracies, and other mistakes that affect data quality.

7.8 GLOSSARY

- **Sampling:** Sampling refers to the process of selecting a subset of individuals or units from a larger population or universe to analyze and derive conclusions about the entire group.
- **Probability Sampling Method:** This approach involves selecting samples based on randomization, ensuring that every element in the population has a known and non-zero chance of being chosen.
- **Simple Random Sampling:** This is the most basic form of random sampling, where each unit has an equal and independent probability of being selected, eliminating bias in the selection process.

- **Stratified Random Sampling:** Also known as proportional or quota random sampling, this technique involves dividing the population into homogeneous subgroups (strata) and then drawing a simple random sample from each subgroup. It is particularly useful for large, diverse populations.
- **Multistage Sampling Method:** This method involves conducting sampling in multiple stages, where each stage consists of clusters of units from the subsequent stage. A suitable random sampling technique is applied at each level. It is especially beneficial for populations spread over extensive geographical areas where a complete sampling frame is unavailable. In multistage sampling, the selection process in the second and further stages is referred to as sub-sampling.

7.9 ANSWERS TO SELF-CHECK EXERCISES

Self-Check Exercise 7.1

Ans. Q1. Refer to Section 7.3.1

Ans. Q2. Refer to Section 7.3.2.2

Ans. Q3. Refer to Section 7.3.2.4

Self-Check Exercise 7.2

Ans. Q1. Refer to Section 7.4

Self-Check Exercise 7.3

Ans. Q1. Refer to Section 7.4

Self-Check Exercise 7.4

Ans. Q1. Refer to Section 7.6.1

Ans. Q2. Refer to Section 7.6.2

7.10 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.

- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House.

7.11 TERMINAL QUESTIONS

Q1. What are the types of Probability or random sampling? What is the Non-probability or non-random sampling?

Q2. Discuss sampling and non-sampling errors?

Unit-08

MEASUREMENT OF CENTRAL TENDENCY: MATHEMATICAL AVERAGE-I

STRUCTURE

8.1 INTRODUCTION

A key goal of statistics is to determine numerical values that effectively describe the fundamental characteristics of a frequency distribution. One of the primary measures used for this purpose is the average. Averages serve to condense large and complex numerical datasets into a single representative value, making interpretation easier. Since the human mind struggles to retain extensive numerical data, it becomes essential to identify a few key constants that summarize the dataset. Averages offer a concise overview, providing insight into the overall pattern of the data. They represent typical values around which other data points tend to cluster, positioned between the highest and lowest observations. This helps in understanding the concentration of values in the central part of the distribution. Due to this property, they are referred to as measures of central tendency. In this unit, we will explore the concept of central tendency, its significance, various types, and methods of measurement.

8.2 LEARNING OBJECTIVES

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

- Define central tendency
- List the objectives of central tendency
- Discuss the uses of central tendency
- Explain the properties of a good average
- Define and calculate Arithmetic Mean

8.3 CENTRAL TENDENCY OR AVERAGE

Averages are also called measures of location since they enable us to locate the position or place of the distribution in question. Averages are statistical constants which enables us to comprehend in a single value the significance of the whole group.

“An average is a typical value that represents all the individual values in a series.”

--Croxton and Cowden

“An average is a single figure that represents the whole group.”

--Clark

In this way we can define an average as a single value within the range of the data that is used to represent all the values in that series. Since an average is somewhere within the range of data, it is sometimes called a measure of central value. An average, is the most typical representative item of the group to which it belongs and which is capable of revealing all important characteristics of that group or distribution.

Self-Check Exercise 8.1

Q1. What is meant by central tendency?

Q2. Define central tendency.

8.4 OBJECTIVES OR FUNCTIONS OF CENTRAL TENDENCY

The objectives of measures of central tendency include:

i) **Summarizing Data**: Measures of central tendency provide a concise summary of the central or typical value within a dataset, allowing researchers, analysts, and decision-makers to grasp the overall characteristics of the data quickly.

ii) **Identifying Typical Values**: By indicating the most common or representative values in the dataset, measures of central tendency help identify typical observations, patterns, and trends within the data.

iii) **Facilitating Comparison**: Central tendency measures enable comparisons between different datasets or subsets of data by providing a common reference point. They help assess similarities, differences, and variations in the central values across various groups or time periods.

iv) **Supporting Decision-Making**: Central tendency measures assist in decision-making processes by providing insights into the central value around which data tends to cluster. This information is valuable for setting benchmarks, establishing targets, and making informed judgments.

v) **Assessing Distribution**: Central tendency measures offer indications of the distributional characteristics of the data, such as symmetry, skewness, or multimodality. They complement measures of dispersion by providing context for understanding the spread of data around the central value.

vi) **Detecting Outliers**: By highlighting the central or typical value within the dataset, central tendency measures help identify outliers or extreme values that may significantly influence the data's overall distribution.

vii) **Interpreting Statistical Analyses**: Central tendency measures serve as essential components of statistical analyses, aiding in the data interpretation of results, hypothesis testing, and drawing meaningful conclusions from research findings.

viii) **Communicating Results**: Central tendency measures provide a clear and concise way to communicate the central value of the data to diverse audiences, including stakeholders, policymakers, and the general public, facilitating understanding and interpretation of statistical information.

Self-Check Exercise 8.2

Q1. What are the objectives of central tendency?

Q2. Explain the various functions of central tendency.

8.5 ESSENTIAL OR PROPERTIES OF A GOOD AVERAGE

An average is a statistical measure used for comparison and analysis. To be effective, it should meet the following criteria:

i) **Precisely Defined** – The average should have a strict definition, eliminating subjective interpretation. A well-defined measure ensures consistency in results when calculated by different individuals.

ii) **Inclusive of All Data Points** – It should take into account all values in the dataset. If any values are excluded, the average may not accurately represent the entire distribution.

iii) **Easily Comprehensible** – The concept of the average should be straightforward and easy to understand. It should not be overly complex or abstract for general use.

iv) **Efficient to Calculate** – The computation should be practical, allowing for quick and accurate calculations without excessive effort.

v) **Stable and Reliable** – The average should remain consistent despite variations in sampling, ensuring reliability in different scenarios.

vi) **Mathematically Adaptable** – It should be suitable for algebraic manipulation, allowing further statistical analysis and applications.

Self-Check Exercise 8.3

Q1. What are essential of a good average?

8.6 TYPES OF AVERAGES

Different methods of measuring Central Tendency provide us with different kinds of averages. These types of averages can be classified as mathematical average and positional average. In mathematical average, algebraic calculation are used to work out the average values. Whereas in positional average, on the basis of position, the average value is calculated. The following are the main types of averages that are commonly used:

8.6.1 Mathematical Average

8.6.1.1 Arithmetic Mean

8.6.1.2 Weighted Mean

8.6.1.3 Geometric Mean

8.6.1.4 Harmonic Mean

8.6.2 Positional Average

8.6.2.1 Median

8.6.2.2 Mode

In this unit, we will discuss about the Arithmetic Mean only. The remaining types of average will be discussed in the next units.

8.6.1.1 Arithmetic Mean

The arithmetic mean of a series is the quotient obtained by dividing the sum of the values by the number of items. In algebraic language, if X_1, X_2, X_3.....X_n are the n values of a variate X, then the Arithmetic Mean (X) is defined by the following formula:

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + X_3 \ldots\ldots\ldots\ldots\ldots + X_n)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum X}{N}$$

(i) Computation of Arithmetic Mean in Individual Series

Example 1: The following are the monthly salaries (in Rs.) of ten employees in an office. Calculate the arithmetic mean

Salary: 250, 275, 265, 280, 400, 490, 670, 890, 1100, 1250

Solution: $\bar{X} = \frac{\sum X}{N}$

$$\bar{X} = \frac{250+275+265+280+400+490+670+890+1100+1250}{10}$$

$$= \frac{5870}{10} = \text{Rs. } 587$$

Short-cut Method: The direct method works well when the dataset is moderate in size, with small and easily manageable values. However, when dealing with a large number of observations or significantly large numerical values, summing all the data points can become time-consuming. To simplify the calculation process, the short-cut method is employed. This method is based on a key property of the arithmetic mean: the algebraic sum of the deviations of individual observations from their mean is always zero. Instead of summing all values directly, deviations from an assumed mean are calculated. The total of these deviations is then divided by the number of observations. The resulting quotient is added to the assumed mean, yielding the arithmetic mean efficiently.

Symbolically, $\bar{X} = A + \frac{\sum dx}{N}$

Where A is assumed mean and dx are deviations= (X-A)

The previous example can be solved by short-cut method as below:

Salary (Rupees)		**Deviations from assumed mean**
S. No.	X	where dx (X – A), A = 400
1	250	–150
2	275	–125
3	265	–135
4	280	–120
5	400	0
6	490	90
7	670	270
8	890	490
9	1100	700
10	1250	850
N = 10		$\sum dx$ = 1870

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

By substituting the values in the formula, we get

$\bar{X}$ = 400 + $\frac{1870}{10}$ = Rs. 587

(ii) **Computation of Arithmetic Mean in Discrete Series**

In a discrete series, the arithmetic mean can be calculated using both the direct method and the shortcut method. The formula for the direct method is:

$\bar{X} = \frac{1}{n}$ ($f_1X_1 + f_2X_2 + f_3X_3$.....+ f_nX_n) = $\frac{\sum(fX)}{N}$

Where the variable values X_1 , X_2, X_3 …….. X_n have frequencies f_1, f_2, f_3…….. f_n and n = $\sum f$

Example 2: The table below presents the distribution of 100 accidents over a week in a specific month. In this month, Fridays and Saturdays occurred five times each, while all other days appeared four times. Determine the average number of accidents per day.

Days:	Sun	Mon	Tue	Wed	Thu	Fri	Sat	**Total**
Number of Accidents:	20	22	10	9	11	8	20 =	**100**

Solution:

Calculation of Number of Accidents per Day

Day	**No. of Accidents**	**No. of Days in Month**	**Total Accidents**
	X	**f**	**fx**
Sunday	20	4	80
Monday	22	4	88
Tuesday	10	4	40
Wednesday	9	4	36
Thursday	11	4	44
Friday	8	5	40
Saturday	20	5	100
	100	$\sum N$ **=30**	$\sum fX$**=428**

$\bar{X}$ = $\frac{\sum(fX)}{N}$ = $\frac{428}{30}$ = 14.27 = 14 accidents per day

The formula for calculating the arithmetic mean using the shortcut method is:

$\bar{X}$ = A + $\frac{\sum(fX)}{N}$ where A is assumed mean, dx = (X-A) and N =$\sum f$

The previous example can be solved using the shortcut method as demonstrated below.

Day	No. of Accidents	dx = (X-A) Where A = 10)	No. of Days in Month	Total Accidents
	X		f	fdx
Sunday	20	10	4	40
Monday	22	12	4	48
Tuesday	10	0	4	0
Wednesday	9	-1	4	-4
Thursday	11	1	4	4
Friday	8	-2	5	-10
Saturday	20	10	5	50
	100		$\sum N$ **=30**	$\sum f\ dx$**=128**

$$\bar{X} = A + \frac{\sum(fX)}{N} = 10 + \frac{128}{30} = 14.27 = 14 \text{ accidents per day}$$

(iii) **Calculation of arithmetic mean for Continuous Series**:

The arithmetic mean can be calculated using the direct method, the short-cut method, or the coding (step deviation) method to simplify computations. Regardless of the method used, determining the mid-values of class intervals in a frequency distribution is essential before computing the arithmetic mean. Once these mid-points are identified, the calculation process follows the same approach as in a discrete series. In the direct method, the arithmetic mean is determined using the formula:

$\bar{X} = \frac{\sum fm}{N}$ where m is mid points of various classes and N =$\sum f$

In the short-cut method, the following formula is applied:

$\bar{X} = A + \frac{\sum fdx}{N}$ where dx = (m-A) and N =$\sum f$

The short-cut method can be further refined using the coding method. In this approach, deviations from the assumed mean are divided by a common factor to reduce their magnitude. The sum of the products of these adjusted deviations and frequencies is then multiplied by the same factor, divided by the total frequency, and added to the assumed mean for the final calculation.

$\bar{X} = A + \frac{\sum fd\prime x}{N}$ X I, where d'x =$\frac{m-A}{i}$, N =$\sum f$ and I = common factor

Example 3: The table below presents the frequency distribution of marks secured by 50 students in a Statistics test.

Marks	Number of Students
0-10	4
10-20	6
20-30	20
30-40	10
40-50	7
50-60	3

Calculate arithmetic mean by;

(i) Direct method,

(ii) Short-cut method, and

(iii) Coding method

Solution:

Calculation of Athematic Mean

X	f	m	fm	$dx = m - A$ (where $A = 25$)	$d'x = \frac{m-A}{i}$ where $i = 10$	fdx	$fd'x$
0—10	4	5	20	– 20	– 2	– 80	– 8
10—20	6	15	90	– 10	– 1	– 60	– 6
20—30	20	25	500	0	0	0	0
30—40	10	35	350	+ 10	+ 1	100	+ 10
40—50	7	45	315	+ 20	+ 2	140	+ 14
50—60	3	55	165	+ 30	+ 3	90	+ 9
	N = 50		$\sum fm = 1440$			$\sum fdx = 190$	$\sum f d' x = +19$

Direct Method:

$$\overline{X} = \frac{\sum fm}{N} = \frac{1440}{50} = 28.8 \text{ marks.}$$

Short-cut Method:

$$\overline{X} = A + \frac{\sum fdx}{N} = 25 + \frac{190}{50} = 28.8 \text{ marks.}$$

Coding Method:

$$\overline{X} = A + \frac{\sum fd'x}{N} \times i = 25 + \frac{19}{50} \times 10 = 25 + 3.8 = 28.8 \text{ marks.}$$

We can observe that answer of average marks i.e. 28.8 is identical by all methods.

8.6.1.1.1 Mathematical Properties of the Arithmetic Mean

(i) Sum of Deviations from the Mean: The total deviation of individual observations from the arithmetic mean always equals zero. Mathematically, this is expressed as:

Symbolically, $\sum(X - \bar{X}) = 0$.

This property establishes the arithmetic mean as the center of gravity, ensuring that the sum of positive deviations balances out the sum of negative deviations.

(ii) Minimization of Squared Deviations: The sum of the squared deviations from the arithmetic mean is the smallest compared to deviations taken from any other value. Symbolically,

$\sum(X - \bar{X})^2$ = smaller than $\sum(X - \text{any other value})^2$.

This property can be verified using sample data.

(iii) Effect of a Constant on the Mean: If each value of a variable XXX is increased, decreased, or multiplied by a constant k, the arithmetic mean will change accordingly in the same manner.

(iv) If we replace each item in the series by the mean, the sum of these substitutions will be equal to the sum of the individual items. This property is used to find out the aggregate values and corrected averages.

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

Or $\sum N = N.\bar{X}$

8.6.1.1.2 Combined Mean

If the arithmetic mean and the number of items in two or more groups are provided, the overall average for these groups can be determined using the following formula:

$$\overline{X}_{12} = \frac{N_1 \overline{X}_1 + N_2 \overline{X}_2}{N_1 + N_2}$$

Where $\bar{X}_{12}$ represents the combined mean of two groups,

$\bar{X}_1$ denotes the arithmetic mean of the first group,

$\bar{X}_2$ signifies the arithmetic mean of the second group

N_1 indicates the number of items in the first group, and

N_2 represents the number of items in the second group.

This concept can be better understood through the following examples.

Example 4:

The mean score of 25 male students in a section is 61, while the mean score of 35 female students in the same section is 58. Determine the overall average score of all 60 students combined.

Solution: We are given the following information,

$$\bar{X}_1 = 61, \qquad N_1 = 25, \qquad \bar{X}_2 = 58, \qquad N_2 = 35$$

$$\text{Apply} \qquad \bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} = \frac{(25 \times 61) + (35 \times 58)}{25 + 35} = 59.25 \text{ marks.}$$

Example 5: The average age of a mixed group of men and women is 30 years. If the mean age of the men in the group is 32 years and that of the women is 27 years, determine the proportion of men and women in the group in percentage terms.

Solution: Let us take group of men as first group and women as second group. Therefore, X1 = 32 years, X_2 = 27 years, and X_{12} = 30 years. In the problem, we are not given the number of men and women. We can assume $N_1 + N_2 = 100$ and therefore, $N_1 = 100 - N_2$.

$$\text{Apply} \qquad \bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$30 = \frac{32N_1 + 27N_2}{100} \qquad (\text{Substitute } N_1 = 100 - N_2)$$

$$30 \times 100 = 32\,(100 - N_2) + 27\,N_2 \qquad \text{or} \qquad 5N_2 = 200$$

$$N_2 = 200/5 = 40\%$$

$$N_1 = (100 - N_2) = (100 - 40) = 60\%$$

Therefore, the proportion of men in the group is 60%, while that of women is 40%.

8.6.1.1.3 Corrected Mean

Example 6: The mean of a dataset containing 100 observations is calculated as 44. However, during computation, two values were mistakenly recorded as 30 and 27 instead of 3 and 72. Determine the corrected mean.

Solution: $\bar{X} = \frac{\sum X}{N}$

$\therefore \quad \sum X = N.\bar{X} = 100 \times 44 = 4400$

$$\text{Corrected} \sum X = \sum X + \text{correct items} - \text{wrong items} = 4400 + 3 + 72 - 30 - 27 = 4418$$

$$\text{Corrected average} = \frac{\text{Corrected} \sum X}{N} = \frac{4418}{100} = 44.18$$

(iv) Calculating the Arithmetic Mean for Open-End Classes

Open-end classes refer to frequency distributions where the lower boundary of the first class and the upper boundary of the last class are not specified. In such cases, determining the arithmetic mean requires making an assumption about these undefined limits. The assumed values typically depend on the class interval of the adjacent defined categories—specifically, the interval following the first class and preceding the last class. For example:

Marks	No. of Students
Below 15	4
15—30	6
30—45	12
45—60	8
Above 60	7

In this instance, since all the defined class intervals are identical, it is assumed that both the first and last classes will have a class interval of 15. As a result, the lower limit of the first class will be set at zero, while the upper limit of the last class will be 75. Consequently, the first class interval will be 0–15, and the last one will be 60–75.

What happens in this case?

Marks	No. of Students
Below 10	4
10—30	7
30—60	10
60—100	8
Above 100	4

In this case, the class interval increases progressively—20 for the second class, 30 for the third, 40 for the fourth, and so on, with a consistent increment of 10. Based on this pattern, it is reasonable to assume that the lower boundary of the first class is 0, while the upper boundary of the final class is 150. For other open-ended class distributions, the first class limit should be established based on the subsequent class interval, whereas the last class limit should be determined using the preceding class interval. When class intervals differ in width, calculating the mean and mode becomes complex and is best avoided. Instead, computing the median is generally the preferred approach.

8.6.1.1.4 Merits of Arithmetic Mean

Arithmetic mean is the most commonly employed measure of central tendency in practice because of the following merits:

i) **Simple to Understand and Easy to Calculate:** The calculation of the arithmetic mean necessitates a basic knowledge of addition, multiplication, and division of numbers. Therefore, even a layman with elementary knowledge can calculate the arithmetic mean. Besides, with the arithmetic mean, it becomes very easy to figure out the value per item or cost per unit, etc.

ii) **Certainty:** An algebraic formula defines the arithmetic mean. Thus, everyone who calculates the average gets the same answer, which ultimately eliminates the chance for deliberate prejudice or personal bias.

iii) **Based on all Items:** The arithmetic mean is calculated by considering all the values. Consequently, it is regarded as being more representative of the distribution.

iv) **Least affected by Fluctuations in Sample:** The arithmetic mean is the least impacted by sampling fluctuations when compared with all other averages. The arithmetic mean provides a suitable basis for comparison in case the number of items in a series is large because abnormalities (errors) in one direction are offset by abnormalities in the other. As a result, the arithmetic mean is considered to be a stable measure.

v) **Convenient Method of Comparison:** Arithmetic Average provides a simple way to compare two or more distributions. For this reason, it is a convenient method for comparison.

vi) **Algebraic Treatment:** Arithmetic mean can be algebraically treated further. It is commonly employed in the computation of several other statistical measures

such as mean deviation, standard deviation, etc., as it can be treated mathematically.

vii) **No Arrangement Required:** The calculation of the arithmetic mean does not require the grouping or arranging of the items. **For example,** in the case of median, the values must be first placed in ascending or descending order, only then the calculation can be done.

8.6.1.1.5 Demerits of Arithmetic Mean

Arithmetic mean matches the majority of the properties for a perfect average, but it has some demerits and needs to be used carefully. Some demerits of arithmetic mean are:

i) **Affected by Extreme Values:** Extreme values have an undue influence on the arithmetic average because it is calculated from all the values in the series, whether they are very small or very large.

ii) **Assumption in the case of Open-end Classes:** It is impossible in the case of open-end classes to calculate the arithmetic mean without making an assumption about the magnitude of the class.

iii) **Absurd Results:** Arithmetic mean can sometimes provide results that appear absurd. **For instance,** if the average number of students (in the school) in a particular class and a particular section comes out to be 15.6, then obviously the result (average) is absurd as children cannot be divided into fractions.

iv) **Not Possible in the case of Qualitative Characteristics:** Arithmetic mean cannot be computed for qualitative data, such as data on IQ, beauty, honesty, etc. The only appropriate average, in this case, is the median.

v) **More Stress on Items of Higher Value:** The arithmetic mean places more emphasis on higher items in a series than it does on lower items. This means that it has an upward bias. If three out of four items have small values and one is quite big, the average will be significantly increased by the big value. However, the opposite is untrue. If a series of four items consists of three items with large values and one item with a small value, the arithmetic average will not be significantly decreased.

vi) **Complete Data Required:** Without all the items in a series, it is impossible to calculate the arithmetic mean. **For instance,** if the values of 99 out of 100 items are known, then the arithmetic average cannot be determined. However, the median and mode averages do not require complete data.

vii) **Calculation by Observation is not Possible:** Arithmetic mean cannot be estimated by simply observing a series, just like in the case of median or mode.

viii) **No Graph Use:** Graphs cannot be used to calculate the arithmetic mean.

ix) **Non-existent Value as Mean:** An arithmetic average might sometimes be a fictional figure that does not exist in the series. 10, 12, 19, and 27 have an arithmetic average of 17. But there is no item in the series that has a value of 17.

Self-Check Exercise 8.4

Q1. What is meant by arithmetic mean?

Q2. What is combined mean?

Q3. Explain the mathematical properties of arithmetic mean.

Q4. What are the merits and demerits of arithmetic mean?

8.7 SUMMARY

Summarisation of the data is a necessary function of any statistical analysis. As a first step in this direction, the huge mass of unwieldy data are summarised in the form of tables and frequency distributions. In order to bring the characteristics of the data into sharp focus, these tables and frequency distributions need to be summarised further. A measure of central tendency or an average is very essential and an important summary measure in any statistical analysis. An average is a single value which can be taken as representative of the whole distribution. In this unit, we have studied meaning of arithmetic mean and its calculation in different types of series.

8.8 GLOSSARY

- **Average:** An average is a single value that falls within the data range and serves as a representative of all values in a given dataset. Because it lies within the data range, it is also referred to as a measure of central tendency.
- **Arithmetic Mean**: Arithmetic Mean is defined as the sum of observations divided by the number of observations.
- **Combined Mean**: When two or more series having different arithmetic means and number of items are combined together, then it is called combined mean. The combined mean of all the series can be calculated using,

$$\overline{X}_{12} = \frac{N_1 \overline{X}_1 + N_2 \overline{X}_2}{N_1 + N_2}$$

8.9 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 8.1

Ans. Q1. Refer to Section 8.3

Ans. Q2. Refer to Section 8.3

Self-Check Exercise 8.2

Ans. Q1. Refer to Section 8.4

Ans. Q2. Refer to Section 8.4

Self-Check Exercise 8.3

Ans. Q1. Refer to Section 8.5

Self-Check Exercise 8.4

Ans. Q1. Refer to Section 8.6.1.1

Ans. Q2. Refer to Section 8.6.1.1.2

Ans. Q3. Refer to Section 8.6.1.1.1

Ans. Q4. Refer to Sections 8.6.1.1.4 and 8.6.1.1.5

8.10 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House

8.11 TERMINAL QUESTIONS

Q1. Define average? What are the functions of an average?

Q2. What are the mathematical properties of arithmetic mean. Explain them with the help of some numerical examples.

Q3. Explain the merits and demerits of arithmetic mean?

Unit-09

MEASUREMENT OF CENTRAL TENDENCY: MATHEMATICAL AVERAGE-II

STRUCTURE

9.1 INTRODUCTION

In the previous unit, we have learnt about the arithmetic mean. The calculation of arithmetic mean in different types of series was discussed with the help of numerical examples. In this unit, we will discuss weighted mean, geometric mean and harmonic mean.

9.2 LEARNING OBJECTIVES

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

- Define and calculate weighted mean
- Discuss and calculate geometric mean
- Describe and calculate harmonic mean

9.3 WEIGHTED MEAN

The term weighted mean refers to the average when different items in the series are assigned different weights based on their corresponding importance. While calculating the simple arithmetic mean, it is assumed that each item in the series has equal importance. However, there are certain cases in which the values of the series observations are not equally important. A simple arithmetic mean will not accurately represent the provided data if all the items are not equally important. Thus, assigning weights to the different items becomes necessary. Different items are assigned different weights based on their relative value. In other words, items that are more significant are given greater weights. The weighted mean is a type of mean that is calculated by multiplying the weight (or probability) associated with a particular event or outcome with its. It can be explained with the help of following example:

Raja Toy Shop offers various toys at different prices: Toy Cars for Rs. 3 each, Toy Locomotives for Rs. 5 each, Toy Aeroplanes for Rs. 7 each, and Toy Double Deckers for Rs. 9 each. If the shop sells one of each type, what would be the average price of the toys?

$\bar{X}$ (Average Price) = $\frac{\sum X}{N}$ = $\frac{24}{4}$ = Rs. 6

In this scenario, each toy holds equal significance since one unit of each type has been sold. This principle is reflected in the calculation of the arithmetic mean, where the price of each toy is considered only once. However, if the shop sells a total of 100 toys—comprising 50 cars, 25 locomotives, 15 airplanes, and 10 double-decker buses—the relative importance of these toys in generating revenue for the dealer varies. The significance of each toy corresponds to the number of units sold, meaning the importance of toy cars is 50, locomotives 25, airplanes 15, and double-decker buses 10.

It may be noted that 50, 25, 15, 10 are the quantities of the various classes of toys sold. These quantities are called as 'weights' in statistical language. Weight is represented by symbol W and ∑W represents the sum of weights.

While determining the average price of toy sold these weights are of great importance and are taken into account to compute weighted mean.

$$\overline{X}_w = \frac{\sum[(W_1X_1)+(W_2X_2)+(W_3X_3)+(W_4X_4)]}{W_1+W_2+W_3+W_4} = \frac{\sum WX}{\sum W}$$

Where, W_1, W_2, W_3, W_4 are weights and X_1, X_2, X_3, X_4 represents the price of 4 varieties of toy.

Hence by substituting the values of W_1, W_2, W_3, W_4 and X_1, X_2, X_3, X_4, we get

$$\bar{X}_w = \frac{(50\times3)+(25\times5)+(15\times7)+(10\times9)}{50+25+15+10}$$

$$\bar{X}_w = \frac{150+125+105+90}{100} = \frac{470}{100} = \text{Rs. } 4.70$$

Example 1: Calculate weighted mean from the following data:

Items	21	36	44	48	53	55
Weight	5	4	3	7	8	3

Solution:

X	W	WX
21	5	105
36	4	144
44	3	132
48	7	336
53	8	424
55	3	165
	$\sum W$= 30	$\sum WX$ =1306

$\bar{X}_w = \frac{\sum WX}{\sum W} = \frac{1306}{30}$ = 45.33

Example 2: A student obtained 60 marks in English, 75 marks in Hindi, 63 marks in Mathematics, 59 marks in Economics, and 55 marks in Statistics. Calculate the weighted mean of the marks if the weights are respectively 2,1,5,5,and 3.

Solution:

Subjects	X	W	WX
English	60	2	120
Hindi	75	1	75
Mathematics	63	5	315
Economics	59	5	295
Statistics	55	3	165
		$\sum W$= 16	$\sum WX$ =970

$\bar{X}_w = \frac{\sum WX}{\sum W} = \frac{970}{16}$ = 60.63

9.3.1 Advantages of Weighted Mean

i) **Accuracy**: The weighted average method takes into account the importance of each value, giving a more accurate representation of the data set. For instance, in calculating the average grade of students in a class, the final exam grade is often given more weight than the other grades in the course, as it is a more accurate measure of the student's knowledge.

ii) **Reflects Importance**: Weighted average is useful when dealing with data that has varying degrees of importance. It is often used in finance, where different types of investments have different risks and returns. The weighted average helps to reflect the importance of each investment and gives an accurate representation of the overall performance of the portfolio.

iii) **Flexibility:** The weighted average method is flexible and can be used in different scenarios, such as calculating the average temperature of a region, where different weather stations have varying levels of accuracy.

9.3.2 Disadvantages of Weighted Mean

i) **Data Availability**: The weighted average method requires the availability of data on the weights of each value, which can be difficult to obtain. For example, if one is to calculate the weighted average of the students' grades in a class, obtaining the weightage of each grade can be a cumbersome task.

ii) **Complexity**: The weighted average method can be complex, especially when dealing with a large amount of data. It requires careful consideration of each value's weightage, which can be time-consuming and prone to errors.

iii) **Misleading**: The weighted average can be misleading if the weights assigned do not accurately reflect the importance of each value. For instance, if a company calculates the weighted average of employee salaries, but assigns a higher weightage to the salaries of the top executives, it may not accurately reflect the average salary of all employees.

The weighted average method is a useful tool for calculating averages when dealing with data that has varying levels of importance. However, it is important to consider the strengths and weaknesses of the method before using it to ensure its effectiveness in handling the data.

Self-Check Exercise 9.1

Q1. What is weighted mean.

Q2. Explain the advantages and disadvantages of weighted mean.

Q3. A student obtained 60 marks in English, 75 marks in Hindi, 63 marks in Mathematics, 59 marks in Economics, and 55 marks in Statistics. Calculate the weighted mean of the marks if the weights are respectively 2,1,5,5,and 3.

9.4 GEOMETRIC MEAN

In general, if we have n numbers (none of them being zero), then the G.M. is defined as

$$G.M. = \sqrt{x_1, x_2, \ldots\ldots\ldots\ldots x_n} = (x_1, x_2, \ldots\ldots\ldots\ldots x_n)^{1/n}$$

In case of a discrete series. If X_1, X_2…….X_n occur f_1, f_2, …….f_n times respectively and N is the total frequency.

G.M. $= \sqrt[n]{X1f1, X2f2, \ldots\ldots\ldots Xnfn}$

For convenience, use of logarithms is made extensively to calculate the nth root. In terms of logarithms

G.M. = AL $\left(\frac{logX1+logX2+ \ldots\ldots\ldots\ldots logXn}{n}\right)$

G.M. = AL $\left(\frac{\sum logx}{N}\right)$ where AL refer to antilog

In case of discrete series, G.M. = AL $\left(\frac{\sum flogx}{N}\right)$

And in case of continuous series, G.M. = AL $\left(\frac{\sum flogm}{N}\right)$

Example 3: Calculate G.M. of the following data: 2, 4, 8

Solution:

G.M. = $\sqrt[3]{2 \times 4 \times 8}$ = $\sqrt[3]{64}$ = 4

In terms of logarithms, the question can be solved as follows:

log 2 = 0.3010, log 4 = 0.6021, and log 8 = 9.9031

Apply the formula:

$$G.M. = AL\frac{\sum \log x}{N} = AL\frac{1.8062}{3} = AL\,(0.60206) = 4$$

Example 4: Calculate geometric mean from the following data

X	5	6	7	8	9	10	11
f	2	4	7	10	9	6	2

Solution: Calculation of G.M.

x	$\log x$	f	$f \log x$
5	0.6990	2	1.3980
6	0.7782	4	3.1128
7	0.8451	7	5.9157
8	0.9031	10	9.0310
9	0.9542	9	8.5878
10	1.0000	6	6.0000
11	1.0414	2	2.0828
		$N = 40$	$\sum f \log x = 36.1281$

$$\text{G.M.} = \text{AL}\left(\frac{\sum f \log x}{N}\right) = \text{AL}\left(\frac{36.1281}{40}\right) = \text{AL}\,(0.9032) = 8.002$$

9.4.1 Weighted G.M.

The weighted G.M. is calculated with the help of the following formula:

$$\text{G.M.} = \sqrt{x_1 w_1, x_2 w_2 \ldots\ldots\ldots\ldots x_n w_n}$$

$$= \frac{\log x_1 \times w_1 + \log x_2 \times w_2 + \ldots\ldots\ldots\ldots\ldots \log x_n \times w_n}{w_1 + w_2 + \ldots\ldots\ldots w_n}$$

$$= \text{AL}\left[\frac{\sum (\log x \times w)}{\sum w}\right]$$

Example 5: Find out weighted G.M. from the following data:

Group	Index Number	Weights
Food	352	48
Fuel	220	10
Cloth	230	8
House Rent	160	12
Misc.	190	15

Solution:

Calculation of Weighted G.M.

Group	Index Number (X)	Weights (W)	Log X	W LogX
Food	352	48	2.5465	122.2320
Fuel	220	10	2.3424	23.4240
Cloth	230	8	2.3617	18.8936
House Rent	160	12	2.2041	26.4492
Misc.	190	15	2.2788	34.1820
		$\sum W = 93$		$\sum WLogx = 225.1808$

$$\text{G. M. (weighted)} = \text{AL}\left[\frac{\sum w \log x}{\sum w}\right] = \text{AL}\,\frac{225.1808}{93} = 263.8$$

Example 6: A machine undergoes depreciation at a rate of 35.5% per year during the first year, followed by 22.5% per year in the second year, and 9.5% per year in the third year. Each depreciation rate is calculated based on the machine's actual value at the beginning of the respective year. Determine the average depreciation rate over the three-year period.

Solution: Average rate of depreciation can be calculated by taking G.M.

Year	***X* (values taking 100 as base)**	**log *X***
I	100 – 35.5 = 64.5	1.8096
II	100 – 22.5 = 77.5	1.8893
III	100 – 9.5 = 90.5	1.9566
		$\sum$ log X = 5.6555

$$\text{Apply}\quad \text{G.M.} = \text{AL}\left[\frac{\sum \log x}{w}\right] = \frac{5.6555}{3} = \text{AL}\,1.8851 = 76.77$$

Example 7: The arithmetic mean and geometric mean of two numbers are given as 10 and 8, respectively. Determine the values of these numbers.

Solution:

Let the two numbers be a and b. Then,

A.M. = $\frac{a+b}{2}$

10 = $\frac{a+b}{2}$ or a+b = 20 ……..(i)

G.M. = $\sqrt[2]{ab}$

8 = $\sqrt[2]{ab}$ or ab = $(8)^2$ = 64……..(ii)

We know,

a-b = $\sqrt{(a+b)^2 - 4ab}$ = $\sqrt{(40)^2 - 4 \times 64}$ = $\sqrt{400 - 256}$ = $\sqrt{144}$ = 12……..(iii)
from equation (i) and (iii)

a + b = 20
a – b = 12

By adding 2a = 32
Or a = 16
and b = 20-16 = 4

9.4.2 Specific Uses of G.M.: The geometric Mean has certain specific uses, some of them are:

(i) It is used in the construction of index numbers,

(ii) It is also helpful in finding out the compound rates of change such as the rate of growth of population in a country.

(iii) It is suitable where the data are expressed in terms of rates, ratios and percentage.

(iv) It is quite useful in computing the average rates of depreciation or appreciation.

(v) It is most suitable when large weights are to be assigned to small items and small weights to large items.

Self-Check Exercise 9.2

Q1. What is meant by geometric mean.

Q2. Explain the specific uses of geometric mean

9.5 HARMONIC MEAN

The harmonic mean is defined as the reciprocals of the average of reciprocals of all items in a series. In other words, the harmonic mean of n observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals.

Calculation of Harmonic Mean

(a) Individual series

If there are n observations X_1, X_2, ... X_n, their harmonic mean is defined as

$$HM = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots\cdots + \frac{1}{X_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

Example 8: Obtain harmonic mean of 15, 18, 23, 25 and 30.

Solution:

$$HM = \frac{5}{\frac{1}{15} + \frac{1}{18} + \frac{1}{23} + \frac{1}{25} + \frac{1}{30}} = \frac{5}{0.239} = 20.92 \text{ Ans.}$$

(b) Ungrouped Frequency Distribution

For ungrouped data, i.e., each X_1, X_2, ... X_n, occur with respective frequency f_1, f_2 .....f_n, where $\sum f_i = N$ is total frequency, the arithmetic mean of the reciprocals of observations is given by

$$\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{X_i}$$

$$\text{Thus,} \quad HM = \frac{N}{\sum \frac{f_i}{X_i}}$$

Example 9: Calculate harmonic mean of the following data:

X	10	11	12	13	14
f	8	5	10	9	6

Solution: **Calculation of Harmonic Mean**

X	10	11	12	13	14	*Total*
Frequency (f)	5	8	10	9	6	38
$f \quad \frac{1}{X}$	0.5000	0.7273	0.8333	0.6923	0.4286	3.1815

$$\therefore \text{HM} = \frac{38}{3.1815} = 11.94$$

(c) Continuous Frequency Distribution

In case of a continuous frequency distribution, the class intervals are given. The mid-values of the first, second ..... nth classes are denoted by X_1, X_2, ...X_n. The formula for the harmonic mean is same, as given in (b) above.

Example 10: Find the harmonic mean of the following distribution:

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	11	21	35	30	22	18

Solution:

Class Intervals	Mid-Values (X)	Frequency (f)	$\frac{1}{x}$	$\frac{f}{x}$
0-10	5	5	0.2000	1.0000
10-20	15	8	0.0667	0.5333
20-30	25	11	0.0400	0.4400
30-40	35	21	0.0286	0.6000
40-50	45	35	0.0222	0.7778
50-60	55	30	0.0182	0.5455
60-70	65	22	0.0154	0.3385
70-80	75	18	0.0133	0.2400
		∑f =150		$\sum \frac{f}{x} = 4.4751$

H.M. = $\frac{150}{4.4751}$ = 33.52 Ans.

9.5.1 Weighted Harmonic Mean

If X_1, X_2,.....X_n are n observations with weights w_1,w_2,.....w_n respectively, their weighted harmonic mean is defined as follows :

$$HM = \frac{\sum w_i}{\sum \frac{w_i}{X_i}}$$

Example 11: A train covers a distance of 50 km at a speed of 40 km/h, followed by 60 km at a speed of 50 km/h, and then 40 km at a speed of 60 km/h. Determine the weighted harmonic mean of the train's speed, using the distances traveled as weights. Additionally, confirm that this harmonic mean serves as a suitable representation of the train's average speed.

Solution: $$HM = \frac{\sum w_i}{\sum \frac{w_i}{X_i}} = \frac{150}{\frac{50}{40}+\frac{60}{50}+\frac{40}{60}} = \frac{150}{1.25+1.20+0.67} \quad \ldots. (1)$$

$$= 48.13 \text{ kms/hour}$$

Verification : $$\text{Average speed} = \frac{\text{Total distance travelled}}{\text{Total time taken}}$$

In Equation (1), the numerator represents the total distance covered by the train. Meanwhile, the denominator indicates the total travel time for a journey of 150 kilometers. This is because the time taken to cover 50 kilometers at a speed of 40 km/h is 50/40. Similarly, the train takes 60/50 and 40/60 hours to travel 60 kilometers and 40 kilometers at speeds of 50 km/h and 60 km/h, respectively. Given these calculations, the weighted harmonic mean serves as the most appropriate measure of average speed in this context.

9.5.2 Advantages of the Harmonic Mean

i) It is a well-defined and precise measure of central tendency.

ii) All data points are taken into account in its calculation.

iii) It assigns lower weights to larger values and higher weights to smaller values.

iv) The harmonic mean allows for further mathematical operations and analysis.

v) It is particularly useful for calculating average rates under specific conditions.

9.5.3 Disadvantages of the Harmonic Mean

i) It is relatively complex to compute and may be difficult to interpret.

ii) The result may not correspond to an actual data point in the given set.

iii) It cannot be determined if any of the observations include a zero value.

iv) If smaller values are assigned proportionally lower weights, the harmonic mean may not effectively represent the dataset.

Self-Check Exercise 9.3

Q1. What is Harmonic Mean.

Q2. Explain the advantages and disadvantages of Harmonic Mean.

Q3. Distinguish between Harmonic Mean and Geometric Mean?

9.6 RELATIONSHIP AMONG AM, GM AND HM

(i) For any two positive number, G.M. = $\sqrt{A.M.\times H.M.}$.

A.M. = $\frac{a+b}{2}$, G.M. = $\sqrt{ab}$, and H.M. = $\frac{2ab}{a+b}$

A.M. × H.M. = $\frac{a+b}{2} \times \frac{2ab}{a+b}$ = ab = $(G.M.)^2$

Hence, the result is proved.

(ii) When all the value of the series differ in size, then *AM* > *GM* > *HM*.

(iii) If all the observations of a variable are same, all the three measures of central tendency coincide, i.e., *AM* = *GM* = *HM*.

Self-Check Exercise 9.4

Q1. Establish the relation between AM, GM and HM?

9.7 SUMMARY

In this unit, we have discuss in detail the weighted mean, geometric mean, and harmonic mean. The specific uses of these averages, their advantages and disadvantages are also discussed in this unit. With the help of some numerical problems we also learnt to calculate these averages. In the next units, we will discuss positional type averages.

9.8 GLOSSARY

- **Average:** An average is a single representative value within a dataset that summarizes the entire series. Since it falls within the data range, it is often referred to as a measure of central tendency.
- **Arithmetic Mean:** The arithmetic mean is calculated by dividing the sum of all observations by the total number of observations.
- **Geometric Mean:** The geometric mean of a set of nnn positive numbers is obtained by taking the nnnth root of their product.
- **Harmonic Mean:** The harmonic mean of n non-zero values is determined by taking the reciprocal of the arithmetic mean of their reciprocals.

9.9 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 9.1

Ans.Q1. Refer to Section 9.3

Ans.Q2. Refer to Sections 9.3.1 and 9.3.2

Ans.Q3. Refer to Section 9.3 (Example 2)

Self-Check Exercise 9.2

Ans.Q1. Refer to Section 9.4

Ans.Q2. Refer to Section 9.4.2

Self-Check Exercise 9.3

Ans.Q1. Refer to Section 9.5

Ans.Q2. Refer to Sections 9.5.2 and 9.5.3

Ans.Q3. Refer to Sections 9.3 and 9.4

Self-Check Exercise 9.4

Ans.Q1. Refer to Section 9.6

9.10 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House

9.11 TERMINAL QUESTIONS

Q1. What is weighted mean. Explain its advantages and disadvantages.

Q2. What is Harmonic Mean. Explain its advantages and disadvantages.

Q3. Distinguish between Harmonic Mean and Geometric Mean?

Q4. Under what circumstances would a geometric mean or a harmonic mean be more appropriate than arithmetic mean?

Unit-10

MEASUREMENT OF CENTRAL TENDENCY: POSITIONAL AVERAGE-I

STRUCTURE

10.1 INTRODUCTION

In the last unit, we started with the measure of central tendency. We have also gone through the mean, harmonic mean and geometric mean. In this unit, we will deal with the meaning of median, its merits and demerits, its calculation in different types of series.

10.2 LEARNING OBJECTIVES

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

- Define Median and calculate the median
- Discuss the merits and demerits of median
- Explain the uses of median

10.3 MEDIAN

The median is that value of the variable which divides the group in two equal parts. One part comprising the values greater than and the other all values less than median. Median of a distribution may be defined as that value of the variable which exceeds and is exceeded by the same number of observation. It is the value such that the number of observations above it is equal to the number of observations below it. Thus we know that the arithmetic mean is based on all items of the distribution, the median is positional average, and that is, it depends upon the position occupied by a value in the frequency distribution.

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

-----Connor

10.3.1 Calculation of Median

(I) In case of Individual Series: The formula used for calculating median in individual series is:

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

Where, N = total number of items

If the number of items is even, then there is no value exactly in the middle of the series. In such a situation the median is arbitrarily taken to be halfway between the two middle items. Symbolically,

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

Example 1: Find the median of the following series:

(i) 8, 4, 8, 3, 4, 8, 6, 5, 10.

(ii) 15, 12, 5, 7, 9, 5, 11, 28.

Solution: (i) Computation of Median

S. No	X	Ascending Order
1	8	3
2	4	4
3	8	4
4	3	5
5	4	6
6	8	8
7	6	8
8	5	8
9	10	10
N= 9		

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

= Size of $\left(\frac{9+1}{2}\right)^{th}$ item = Size of 5^{th} item = 6

Solution: (ii) Computation of Median

S. No	X	Ascending Order
1	15	5
2	12	5
3	5	7
4	7	9
5	9	11
6	5	12
7	11	15
8	28	28
N=8		

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

= Size of $\left(\frac{8+1}{2}\right)^{th}$ item =

= $\frac{Size\ of\ 4th\ item + Size\ of\ 5th\ item}{2}$ = $\frac{9+11}{2}$ = 10

(II) In case of Discrete series: In a discrete series, medium is computed in the following manner:

(i) Arrange the given variable data in ascending or descending order.

(ii) Find cumulative frequencies.

(iii) Apply median = Size of $\left(\frac{N+1}{2}\right)^{th}$ item

(iv) Locate median according to the size i.e., variable corresponding to the size or for next cumulative frequency.

Example 2: Following are the number of rooms in the houses of a particular locality. Find median of the data:

No. of Rooms:	3	4	5	6	7	8
No. of Houses:	38	654	311	42	12	2

Solution: Computation of Median

No. of Rooms X	No. of Houses f	Cumulative Frequency Cf
3	38	38
4	654	692
5	311	1003
6	42	1045
7	12	1057
8	2	1059

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

= Size of $\left(\frac{1059+1}{2}\right)^{th}$ item = 530^{th} item

Median lies in the cumulative frequency of 692 and the value corresponding to this is 4. Therefore, Median = 4 rooms.

(III) In case of Continuous Series: Median is computed in continuous series in the following manner:

(i) Arrange the given variable data in ascending or descending order.

(ii) If inclusive series is given, it must be converted into exclusive series to find real class intervals.

(iii) Find cumulative frequencies.

(iv) Apply Median = size of $\frac{N}{2}$ th item to ascertain median class.

(v) Apply formula of interpolation to ascertain the value of median.

$$\text{Median} = l_1 + \frac{\frac{N}{2} - cf_0}{f} \times (l_1 - l_2) \qquad \text{or} \qquad \text{Median} = l_2 - \frac{\frac{N}{2} - cf_0}{f} \times (l_2 - l_1)$$

where, l_1 refers to lower limit of median class,

l_2 refers to higher limit of median class,

cf_0 refers cumulative frequency of previous to median class,

f refers to frequency of median class,

Example 3: The following table gives you the distribution of marks secured by some students in an examination:

Marks	**No. of Students**
11—20	42
21—30	48
31—40	120
41—50	84
51—60	48
61—70	36
71—80	31

Find the median marks.

Solution: Calculation of Median Marks

Marks (*x*)	No. of Students (*f*)	*cf*
11—20	42	42
21—30	38	80
31—40	120	200
41—50	84	284
51—60	48	332
61—70	36	368
71—80	31	399

Median = size of $\frac{N}{2}$ th item = size of $\frac{399}{2}$th item = 199.5th item.

Which lies in (31—40) group, therefore the median class is 30.5—40.5.

Applying the formula of interpolation.

$$\text{Median} = l_1 + \frac{\frac{\text{N}}{2} - cf_0}{f} \times (l_1 - l_2)$$

$$= 30.5 + \frac{199.5 - 80}{120} \times (10) = 30.5 + \frac{119.5}{12} = 40.46 \text{ marks.}$$

Calculation of Missing Frequencies

Example: In the frequency distribution of 100 families given below; the number of families corresponding to expenditure groups 20—40 and 60—80 are missing from the table. However the median is known to be 50. Find out the missing frequencies.

Expenditure	0-20	20-40	40-60	60-80	80-100
No. of families	14	?	27	?	15

Solution: We shall assume the missing frequencies for the classes 20-40 to be x and 60-80 to y

Expenditure (Rs.)	No. of Families	C.f.
0—20	14	14
20—40	x	14 + x
40—60	27	14 + 27 + x
60—80	y	41 + x + y
80—100	15	41 + 15 + x + y

N = 100 = 56 + x + y

From the table, we have N = ∑F = 56 + x + y =100.

Therefore, $x + y = 100 - 56$

x +y = 44 ……(i)

Hence we are given M = 50, therefore it lies in 40-60 class interval

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

$50 = 40 + \frac{\frac{100}{2} - (14 + x)}{27} \times 20$

$50\text{-}40 = \frac{50 - 14 - x}{27} \times 20$

$10 = \frac{36 - x}{27} \times 20$

270 = 720- 20x

20x = 720-270

X = 22.5

By substituting the value of x in the equation (i)

Y = 44-22.5 = 21.5

Hence the missing frequencies for the class 20-40 is 22.5 and 60-80 is 21.5.

10.3.2 Properties of Median

1. It is a positional average.

2. It can be shown that the sum of absolute deviations is minimum when taken from median. This property implies that median is centrally located.

10.3.3 Merits and Demerits of Median

(a) Merits

1. It is easy to understand and easy to calculate, especially in series of individual observations and ungrouped frequency distributions. In such cases it can even be located by inspection.
2. Median can be determined even when class intervals have open ends or not of equal width.
3. It is not much affected by extreme observations. It is also independent of range or dispersion of the data.
4. Median can also be located graphically.
5. It is centrally located measure of average since the sum of absolute deviation is minimum when taken from median.
6. It is the only suitable average when data are qualitative and it is possible to rank various items according to qualitative characteristics.
7. Median conveys the idea of a typical observation.

(b) Demerits

1. In case of individual observations, the process of location of median requires their arrangement in the order of magnitude which may be a cumbersome task, particularly when the number of observations is very large.
2. It, being a positional average, is not capable of being treated algebraically.
3. In case of individual observations, when the number of observations is even, the median is estimated by taking mean of the two middle-most observations, which is not an actual observation of the given data.
4. It is not based on the magnitudes of all the observations. There may be a situation where different sets of observations give same value of median.

For example, the following two different sets of observations have median equal to 30.

Set I : 10, 20, 30, 40, 50 and Set II : 15, 25, 30, 60, 90.

5. In comparison to arithmetic mean, it is much affected by the fluctuations of sampling.
6. The formula for the computation of median, in case of grouped frequency distribution, is based on the assumption that the observations in the median class are uniformly distributed. This assumption is rarely met in practice.
7. Since it is not possible to define weighted median like weighted arithmetic mean, this average is not suitable when different items are of unequal importance.

10.3.4 Uses of Median

1. It is an appropriate measure of central tendency when the characteristics are not measurable but different items are capable of being ranked.
2. Median is used to convey the idea of a typical observation of the given data.

3. Median is the most suitable measure of central tendency when the frequency distribution is skewed. For example, income distribution of the people is generally positively skewed and median is the most suitable measure of average in this case.
4. Median is often computed when quick estimates of average are desired.
5. When the given data has class intervals with open ends, median is preferred as a measure of central tendency since it is not possible to calculate mean in this case.

Self-Check Exercise 10.1

Q1. What is meant by median.

Q2. What are merits and demerits od median

Q3. The following table gives you the distribution of marks secured by some students in an examination. Find the median.

Marks	11-20	21-30	31-40	41-50	51-60	61-70	71-80
No of Students	42	48	120	84	48	36	31

10.4 SUMMARY

In this unit, we have gone through the meaning of median and its calculation in different types of series. The merits and demerits of median along with its important uses have also been discussed in this unit.

10.5 GLOSSARY

- **Average:** An average is a representative value within a dataset that summarizes all the values in the series. Since it falls within the data range, it is also referred to as a measure of central tendency.
- **Median:** The median is the value that divides a dataset into two equal halves, with one half containing values greater than the median and the other half containing values less than the median.
- **Cumulative Frequency:** Cumulative frequency refers to the running total of frequencies across different class intervals in a dataset.

10.6 ANSWERS TO SELF CHECK EXERCISES

Self-Check Exercise 10.1

Ans. Q1. Refer to Section10.3

Ans. Q2. Refer to Section10.3.3

Ans. Q3. Refer to Section10.3 (Example 3)

10.7 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.

- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House

10.8 TERMINAL QUESTIONS

Q1. What is median. Explain its merits and demerits.

Q2. Find median from the series given below:

Marks (Less than)	5	10	15	20	25	30	35	40	45	50
No. of Students	5	13	28	53	83	105	123	135	142	145

Q3. Calculate median from the following table:

Wages (More than)	30	40	50	60	70	80	90
No. of Workers	58	46	40	31	16	5	0

Q4. Calculate median from the following data:

CI	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
f	5	8	5	22	20	25	19	25	6	5

Unit-11

MEASUREMENT OF CENTRAL TENDENCY: POSITIONAL AVERAGE-II

STRUCTURE

11.1 INTRODUCTION

In the previous unit, we explored the concept of the median, which divides a distribution into two equal halves. However, a distribution can also be divided into more than two equal parts. The values that accomplish this are referred to as partition values or fractiles. This unit will cover some of the key partition values in detail.

11.2 LEARNING OBJECTIVES

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

- Define partition values
- Explain and calculate lower and upper quartiles,
- Explain and calculate different types of deciles,
- Discuss and calculate different types of parcentiles

11.3 PARTITION VALUES

Just as median divides the series into two equal parts, there are other useful measures which divides the series into four, ten, or hundred equal parts. They are called quartiles, deciles, and percentiles**.**

11.3.1 Quartiles

Quartiles are specific values that divide a dataset into four equal sections. Since three values are required to achieve this division, there are three quartiles: Q1 (the first quartile), Q2 (the second quartile), and Q3 (the third quartile). In the case of a discrete distribution, the first quartile (Q1) is the value at which at least 25% of the observations are either equal to or less than it, while at least 75% of the observations are equal to or greater than it.

For a continuous or grouped frequency distribution, Q1 represents the point where the histogram's area to the left of Q1 accounts for 25% of the total area, and the area to its right makes up the remaining 75%. The formula for calculating Q1 is derived from the median formula by making necessary adjustments. Once the first quartile class is identified, Q1 is determined using the following formula:

$$Q_1 = L_{Q1} + \frac{\left(\frac{N}{4} - C\right)}{fQ1} \times h$$

Here, L_{Q1} is lower limit of the first quartile class, h is its width, f_{Q1} is its frequency and C is cumulative frequency of classes preceding the first quartile class.

The second quartile (Q2) corresponds to the median of the dataset. The third quartile (Q3) is determined similarly to Q1.

For a discrete distribution, Q3 is defined as the value where at least 75% of the observations are less than or equal to it, and at least 25% of the observations are greater than or equal to it. In the case of a grouped frequency distribution, Q3 is the value where the histogram's area to the left of Q3 constitutes 75% of the total area, while the area to its right accounts for 25%. The formula for calculating Q3 is as follows:

$$Q_3 = L_{Q3} + \frac{\left(\frac{3N}{4} - C\right)}{fQ3} \times h$$, Where the symbols have their usual meaning.

11.3.2 Deciles

Deciles partition a distribution into ten equal segments, resulting in nine deciles labeled as D_1, D_2, ..., D_9. In the case of a discrete distribution, the ith decile (D_i) represents the value of the variable for which at least (10i)% of the observations are less than or equal to it, while at least (100 - 10i)% of the observations are greater than or equal to it, where i ranges from 1 to 9.

For a continuous or grouped frequency distribution, D_i corresponds to the value of the variable at which the cumulative area under the histogram to the left of its ordinate equals (10i)%, while the area to the right accounts for (100 - 10i)%. The formula for calculating the ith decile is as follows:

$$D_1 = L_{D1} + \frac{\left(\frac{iN}{10} - C\right)}{fD1} \times h \qquad (i = 1,2,\ldots\ldots\ldots\ldots.9)$$

11.3.4 Percentiles

Percentiles partition a distribution into 100 equal segments, resulting in 99 percentiles labeled as P_1, P_2, ... P_{25}, ... P_{40}, ... P_{60}, ... P_{99}. In the case of a discrete distribution, the kth percentile (Pk) represents the value of the variable for which at least k% of the data points are less than or equal to it, while at least (100 - k)% of the data points are greater than or equal to it.

For a grouped frequency distribution, P□ is the value of the variable such that the proportion of the histogram's area to the left of P□ corresponds to k%, while the area to the right accounts for (100 - k)%. The formula for calculating the kth percentile is given by:

$$P_k = L_{pk} + \frac{\left(\frac{kN}{100} - C\right)}{fpk} \times h \qquad (k = 1,2,\ldots\ldots\ldots\ldots 99)$$

Remarks:

(i) It is important to note that P_{25} corresponds to Q_1, P_{50} is equivalent to D_5, Q_2, and the median (M_d), while P_{75} aligns with Q_3. Similarly, P_{10} matches D_1, and P_{20} corresponds to D_2, and so on.

(ii) Expanding on this, when a distribution is divided into five equal parts, the partition values are referred to as Quintiles, whereas dividing it into eight equal parts results in Octiles.

(iii) The formulas for different partition values in a grouped frequency distribution, as discussed earlier, are derived using cumulative frequencies of the 'less than' type. The equivalent formulas for the 'greater than' type cumulative frequencies can be formulated similarly, as shown below.

$$Q_1 = U_{Q_1} - \frac{\left(\frac{3N}{4} - C\right)}{f_{Q_1}} \times h, \quad Q_3 = U_{Q_3} - \frac{\left(\frac{N}{4} - C\right)}{f_{Q_3}} \times h$$

$$D_i = U_{D_i} - \frac{\left[\left(N - \frac{iN}{10}\right) - C\right]}{f_{D_i}} \times h, \qquad P_k = U_{P_k} - \frac{\left[\left(N - \frac{kN}{100}\right) - C\right]}{f_{P_k}} \times h$$

Here U_{Q1}, U_{Q3}, U_{Di}, U_{PK} are the upper limits of the corresponding classes and C denotes the greater than type cumulative frequencies.

11.3.4 Calculation of Partition Values

For Individual and Discrete Series	For Continuous Series	Formula to be used in Continuous Series
Q_1 = Size of $\frac{N+1}{4}$ th item	Q_1 = Size of $\frac{N}{4}$ th item	$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times i$
Q_3= Size of $\frac{3(N+1)}{4}$ th item	Q_3= Size of $\frac{3N}{4}$ th item	$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times i$
D_1 = Size of $\frac{N+1}{10}$ th item	D_1 = Size of $\frac{N}{10}$ th item	$D_1 = L + \frac{\frac{N}{10} - cf}{f} \times i$
D_9 = Size of $\frac{9(N+1)}{10}$ th item	D_9 = Size of $\frac{9N}{10}$ th item	$D_9 = L + \frac{\frac{9N}{10} - cf}{f} \times i$
P_1 = Size of $\frac{N+1}{100}$ th item	P_1 = Size of $\frac{N}{100}$ th item	$P_1 = L + \frac{\frac{N}{100} - cf}{f} \times i$
P_{99} = Size of $\frac{100(N+1)}{100}$ th item	P_{99} = Size of $\frac{100N}{100}$ th item	$P_{99} = L + \frac{\frac{99N}{10} - cf}{f} \times i$

(i) In case of Individual Series

Example 1: From the following data, calculate Q_1, Q_3, D_5, and P_{25}

21, 15, 40, 30, 26, 45, 50, 54, 60, 65, 70

Solution:

S No	X	Ascending Order
1	21	15
2	15	21
3	40	26
4	30	30
5	26	40
6	45	45
7	50	50
8	54	54
9	60	60
10	65	65
11	70	70
N=11		

Q_1 = Size of $\frac{N+1}{4}$ th item

= Size of $\frac{11+1}{4}$ th item = Size of 3rd item = 26

Q_3 = Size of $\frac{3(N+1)}{4}$ th item

= Size of $\frac{3(11+1)}{4}$ th item = Size of 9th item = 60

D_5 = Size of $\frac{5(N+1)}{10}$ th item

= Size of $\frac{5(11+1)}{10}$ th item = Size of 6th item = 45

P_{25} = Size of $\frac{25(N+1)}{100}$ th item

= Size of $\frac{25(11+1)}{100}$ th item = Size of 3rd item = 26

(ii) In case of Discrete Series:

Example 2: Calculate Q_1, Q_3, D_7, and P_{45} from the following data

X	2	4	6	8	10	12	14	16	18
f	4	3	5	8	10	11	8	6	5

Solution:

X	f	Cumulative frequency
2	4	4
4	3	7
6	5	12
8	8	20
10	10	30
12	11	41
14	8	49
16	6	55
18	5	60
	N=60	

Q_1 = Size of $\frac{N+1}{4}$ th item

= Size of $\frac{60+1}{4}$ th item = Size of 15.25^{th} item = 8

Q_3 = Size of $\frac{3(N+1)}{4}$ th item

= Size of $\frac{3(60+1)}{4}$ th item = Size of 45.25^{th} item = 14

D_7 = Size of $\frac{7(N+1)}{10}$ th item

= Size of $\frac{7(60+1)}{10}$ th item = Size of 42.7^{th} item = 14

P_{45} = Size of $\frac{45(N+1)}{100}$ th item

= Size of $\frac{45(60+1)}{100}$ th item = Size of 27.45^{th} item = 10

(iii) In case of Continuous Series:

Example 3: Calculate Q_1, Q_3, D_6, and P_{68} from the following data

X	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
f	4	6	10	15	9	6	4	3	2

Solution:

X	f	Cf
0-10	4	4
10-20	6	10
20-30	10	20
30-40	15	35
40-50	9	44
50-60	6	50
60-70	5	55
70-80	3	58
80-90	2	60
	N = 60	

Q_1 = Size of $\frac{N}{4}$ th item

= Size of $\frac{60}{4}$ th item = Size of 15^{th} item

Therefore, **Q_1** lies in 20-30 class interval

$\mathbf{Q_1}$ = L + $\frac{\frac{N}{4}-cf}{f} \times i$

= 20 + $\frac{\frac{60}{4}-10}{10} \times 10$ = 20 + $\frac{15-10}{10} \times 10$ = 20+5 = **25 Ans.**

$\mathbf{Q_3}$ = Size of $\frac{3N}{4}$ th item

= Size of $\frac{180}{4}$ th item = Size of 45th item

Therefore, $\mathbf{Q_3}$ lies in 50-60 class interval

$\mathbf{Q_3}$ = L + $\frac{\frac{3N}{4}-cf}{f} \times i$

= 50 + $\frac{\frac{180}{4}-44}{6} \times 10$ = 50 + $\frac{45-44}{6} \times 10$ = 50+1.667 = **51.667 Ans.**

$\mathbf{D_6}$ = Size of $\frac{6N}{10}$ th item

= Size of $\frac{360}{10}$ th item = Size of 36th item

Therefore, $\mathbf{D_6}$ lies in 40-50 class interval

$\mathbf{D_6}$ = L + $\frac{\frac{6N}{10}-cf}{f} \times i$

= 40 + $\frac{\frac{360}{10}-35}{9} \times 10$ = 40 + $\frac{36-35}{9} \times 10$ = 40+1.111 = **41.111 Ans.**

$\mathbf{P_{68}}$ = Size of $\frac{68N}{100}$ th item

= Size of $\frac{4080}{100}$ th item = Size of 40.80th item

Therefore, $\mathbf{P_{68}}$ lies in 40-50 class interval

$\mathbf{P_{68}}$ = L + $\frac{\frac{68N}{100}-cf}{f} \times i$

= 40 + $\frac{\frac{4080}{100}-35}{9} \times 10$ = 40 + $\frac{40.80-35}{9} \times 10$ = 40+6.444 = **46.444 Ans.**

Self-Check Exercise 11.1

Q1. What do you mean by partition values?

Q2. What is meant by quartiles, deciles, and percentiles.

Q3. Following data show marks of students in first class test. Calculate Q_1, Q_3, D_1, and P_{95} from the following data.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Students	11	18	25	28	30	33	22	15	12	10

11.4 SUMMARY

The Partition Values are the measures used to divide the total number of observations from a distribution into a certain number of equal parts. Quartiles, Deciles, and Percentiles are some of the most often used partition values. For the purposes of summary, we have the meaning of partition values; these concepts have a wide range of applications in our day to day data analysis tasks.

11.5 GLOSSARY

- **Average**: An average is a single value that falls within the data range and serves as a representative figure for all values in a dataset. Since it lies within the range of observations, it is often referred to as a measure of central tendency.
- **Median**: The median is the value that divides a dataset into two equal halves. One half consists of values greater than the median, while the other half contains values lower than it.
- **Quartiles**: Quartiles are values that split a dataset into four equal sections. Since three values are required to create these divisions, there are three quartiles—Q1 (first quartile), Q2 (second quartile), and Q3 (third quartile).
- **Deciles**: Deciles divide a dataset into ten equal segments. There are nine deciles in total, represented as D1, D2, ..., D9.
- **Percentiles**: Percentiles divide a dataset into 100 equal parts. There are 99 percentiles, labeled as P1, P2, ..., P25, ..., P40, ..., P60, ..., P99.

11.6 ANSWERS TO SELF-CHECK EXERCISES

Self-Check Exercise 11.1

Ans. Q1. Refer to Section 11.3

Ans. Q2. Refer to Sections 11.3.1, 11.3.2 and 11.3.3

Ans. Q3. Q1=28.8, Q3=62.73, D1=15.22, D9=81.33

11.7 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House, New Delhi.
- Jain, T.R. and Aggarwal, S.C. (2022). Business Statistics. V.K Global Publications Pvt. Ltd. New Delhi.

11.8 TERMINAL QUESTIONS

Q1. What do you mean by partition values?

Q2. Write a detailed note on calculation of quartiles, deciles, and percentiles.

Q3: Calculate Median, Q_1, Q_3, D_4, D_7, P_{15}, P_{60} and P_{90} from the following data:

Daily Profit	75	76	77	78	79	80	81	82	83	84	85
No of Shops	15	20	32	35	33	22	20	10	8	3	2

Q4. Calculate Q_1, Q_3, D_7, and P_{65} from the following data.

Marks (Less than)	5	10	15	20	25	30	35	40	45	50
No. of Students	5	13	28	53	83	105	123	135	142	145

Q5. Calculate Q_1, Q_3, D_6, and P_{80} from the following data.

Wages (More than)	30	40	50	60	70	80	90
No. of Workers	58	46	40	31	16	5	0

Unit-12

MEASUREMENT OF CENTRAL TENDENCY: POSITIONAL AVERAGE-III

STRUCTURE

12.1 INTRODUCTION

In statistics, the mode is the value that is repeatedly occurring in a given set. We can also say that the value or number in a data set, which has a high frequency or appears more frequently, is called mode or modal value. It is one of the three measures of central tendency, apart from mean and median. For example, the mode of the set {3, 7, 8, 8, 9}, is 8. Therefore, for a finite number of observations, we can easily find the mode. A set of values may have one mode or more than one mode or no mode at all. In this unit, we will understand the meaning of mode in statistics, formula for the calculation of mode value and solve the numerical problems in detail.

12.2 LEARNING OBJECTIVES

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

- Define Mode
- Calculate Mode in the discrete series
- Calculate mode in the continuous series
- List the merits and demerits of Mode

12.3 MODE

The mode is the value in a dataset that appears most frequently. It represents the most common or typical observation, making it the most "popular" or frequently occurring value.

It is defined by Zizek as “the value occurring most frequently in series of items and around which the other items are distributed most densely.”

In the words of Croxton and Cowden, “the mode of a distribution is the value at the point where the items tend to be most heavily concentrated.”

According to A.M. Tuttle, “Mode is the value which has the greater frequency density in its immediate neighbourhood.”

For a set of individual observations, the mode refers to the value that appears most frequently in the dataset. It is represented by the letter Z.

12.3.1 Calculation of Mode

(i) in case of Individual Series: In case of individual series, mode can be identify as a value that occur most frequently in a series**.**

Example 1: Calculate mode from the following data of the marks obtained by students

10, 27, 24, 12, 27, 27, 20, 18, 15, 30

Solution:

Marks	No. of Students
10	1
12	1
15	1
18	1
20	1
24	1
27	3
30	1

Mode is 27 marks because it occurs most of the times.

(ii) In case of Discrete Series: For calculating mode in discrete series, the following two methods are used:

(a) Inspection Method: in this method, the value of mode is determined by inspecting the series. The value whose frequency is maximum is mode.

Example 2: find mode from the following data:

Shoe Size	1	2	3	4	5	6	7
frequency	4	5	13	6	12	8	6

Solution: By inspection, the modal size of the series is 3 as it has the maximum frequency (13).

(b) Grouping Method: in some cases, it is possible that the value having the highest frequency may not be the modal value. This will specially be so where the difference between the maximum frequency and the frequency proceeding or succeeding is very small and items are heavily concentrated on either side. Under the grouping method, modal value is determined by preparing two tables-(i) Grouping Table and (ii) Analysis Table. It can be explained with the help of above example.

Example 3: find mode from the following data:

Shoe Size	1	2	3	4	5	6	7
frequency	4	5	13	6	12	8	6

Solution: By observation, the modal value appears to be 3, as it has the highest frequency. However, this method is not entirely reliable since the mode is influenced not only by the frequency of a single class but also by the frequencies of adjacent classes. In such situations, the Grouping and Analysis table method is used.

While the mode is initially identified as 3, its actual value might be 5 because the neighboring frequencies for size 5 are higher than those for size 3. This impact of adjacent frequencies can be analyzed effectively using the Grouping and Analysis table technique.

Grouping Table

Shoe Size	Column I	Column II	Column III	Column IV	Column V	Column VI
1	4	} 9				
2	5		} (18)	] 22		
3	(13)	} 19			] 24	
4	6		} (18)			] (31)
5	12	} (20)		] (26)		
6	8		} 14		} (26)	
7	6					

This grouping table has seven columns. Column I shows the frequency and circle the highest frequency. In Column II , frequencies are grouped in two's, starting with the first two frequencies of the series. In Column III, first frequency is left out and the remaining are grouped in two's. In Column fourth, frequencies are grouped in three's starting the first three frequencies. In Column fifth, leave the first frequency and group

the remaining in three's. in Column Sixth, leave the two frequency and group the remaining in three's. Circle the highest frequency in each column. The six columns are to serve as the basis for the preparation of Analysis Table.

For the preparation of the Analysis Table, enter the tick mark (√) into the the column the highest frequencies marks in the Grouping Table . Take the total of each column to find out the value repeated maximum number of times. This value against which the total is the highest is the mode.

Analysis Table

Shoes Size	Column I	Column II	Column III	Column IV	Column V	Column VI	Total
1							**0**
2			√				**1**
3	√		√			√	**3**
4			√	√		√	**3**
5		√	√	√	√	√	**5**
6		√		√	√		**3**
7					√		**1**

Item 5 appears most frequently, making it the mode. However, upon inspection, we initially identified 3 as the mode.

(ii) **For a Continuous Series**

In a continuous series, determining the mode involves an additional step. After identifying the modal class through inspection or the grouping technique, the mode is calculated using the interpolation formula:

$$\mathbf{Z = L + \frac{f1-f0}{2f1-f0-f2} \times i}$$

Z = Value of Mode.

L = lower limit of the class, where mode lies.

f_0 = frequency of the class proceeding the modal class.

f_1 = frequency of the class, where mode lies.

f_2 = frequency of the class succeeding the modal class.

Example 4: Calculate mode of the following frequency distribution:

Variable	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	5	10	15	14	10	5	3

Solution:

Grouping Table

Variable	Column I	Column II	Column III	Column IV	Column V	Column VI
0-10	5	15				
10-20	10		25	30		
20-30	15	29			39	39
30-40	14		24	29		
40-50	10	15			18	
50-60	5		8			
60-70	3					

Analysis Table

Variable	Column I	Column II	Column III	Column IV	Column V	Column VI	Total
0-10				√			1
10-20			√	√	√		3
20-30	√	√	√	√	√	√	6
30-40		√			√	√	3
40-50						√	1
50-60							0
60-70							0

Therefore, Z lies in class interval 20-30 because it has the highest total. Mode is calculated by using following formula:

$$\mathbf{Z} = \mathbf{L} + \frac{f1-f0}{2f1-f0-f2} \times i$$

$$= 20 + \frac{15-10}{2(15)-10-14} \times 10$$

$$= 20 + \frac{5}{6} \times 10 = 20+8.333 = 28.333$$

(iv) Calculation of Mode where it is ill defined:

This formula does not apply when a series or distribution contains multiple modal values. For example, if two or more items share the highest frequency, the series is classified as bimodal or multimodal. In such situations, the mode is considered ill-defined, and the following formula is used instead.

Mode = 3 Median – 2 Mean.

Example 5: Calculate mode of the following frequency data:

Class Interval	Frequency
10—20	5
20—30	9
30—40	13
40—50	21
50—60	20
60—70	15
70—80	8
80—90	3

Solution: First of all, ascertain the modal group with the help of process of grouping.

Grouping Table

Class Interval	Column I	Column II	Column III	Column IV	Column V	Column VI
10-20	5	14		27		
20-30	9		22		43	
30-40	13	34				54
40-50	21		41	56		
50-60	20	35			43	
60-70	15		23			26
70-80	8	11				
80-90	3					

Analysis Table

CI	Column I	Column II	Column III	Column IV	Column V	Column VI	Total
10-20							0
20-30					√		1
30-40					√	√	2
40-50	√		√	√	√	√	5
50-60		√	√	√	√	√	5
60-70		√		√	√		2
70-80					√		1
80-90							0

There are two groups which occur equal number of items. They are 40-50 and 50-60. Therefore, we will apply the following formula: **Mode = 3 Median – 2 Mean.**

For this purpose, the value of mean and median are required to be computed.

For Arithmetic Mean:

Class Interval	f	Mid Value	d'x= $\frac{X-45}{10}$	fd'x
10-20	5	15	-3	-15
20-30	9	25	-2	-18
30-40	13	35	-1	-13
40-50	21	45 = A	0	0
50-60	20	55	1	+20
60-70	15	65	2	+30
70-80	8	75	3	+24
80-90	3	85	4	+12
	$\sum f = 94$			$\sum fd'x = 40$

$\bar{X}$ = A + $\frac{\sum fd'x}{\sum f} \times i$

= 45 + $\frac{40}{94} \times 10$ = 45+4.2 = 49.2

For Median:

Class Interval	f	Cf
10-20	5	5
20-30	9	14
30-40	13	27
40-50	21	48
50-60	20	68
60-70	15	83
70-80	8	91
80-90	3	94
	$\sum f = 94$	

Median = Size of $\left(\frac{N}{2}\right)^{th}$ item

= Size of $\left(\frac{94}{2}\right)^{th}$ item

= Size of 47^{th} item

Therefore, Median lies in 40-50 class interval

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

$$= 40 + \frac{\frac{94}{2} - 27}{21} \times 10 = 40 + \frac{47-27}{21} \times 10 = 40 + 9.5 = 49.5$$

Mode = 3 Median – 2 Mean

= 3(49.5) – 5(49.2) = 148.5 – 98.4 = **50.1**

Self-Check Exercise 12.1

Q1. Define Mode? Write the formula to calculate Mode in continuous series.

Q2. Calculate mode from the following data of the marks obtained by students

10, 27, 24, 12, 27, 27, 20, 18, 15, 30

Q3. Calculate mode of the following frequency distribution:

CI	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	5	10	15	14	10	5	3

12.4 MERITS AND DEMERITS OF MODE

Merits

i) It is easy to understand and easy to calculate. In many cases it can be located just by inspection.
ii) It can be located in situations where the variable is not measurable but categorization or ranking of observations is possible.
iii) Like mean or median, it is not affected by extreme observations. It can be calculated even if these extreme observations are not known.
iv) It can be determined even if the distribution has open end classes.
v) It can be located even when the class intervals are of unequal width provided that the width of modal and that of its preceding and following classes are equal.
vi) It is a value around which there is more concentration of observations and hence the best representative of the data.

Demerits

i) It is not based on all the observations.
ii) It is not capable of further mathematical treatment.
iii) In certain cases mode is not rigidly defined and hence, the important requisite of a good measure of central tendency is not satisfied.
iv) It is much affected by the fluctuations of sampling.
v) It is not easy to calculate unless the number of observations is sufficiently large and reveal a marked tendency of concentration around a particular value.
vi) It is not suitable when different items of the data are of unequal importance.
vii) It is an unstable average because, mode of a distribution, depends upon the choice of width of class intervals.

Self-Check Exercise 12.2

Q1.. What are the merits and demerits of Mode?

12.5 RELATION BETWEEN MEAN, MEDIAN AND MODE

The connection between these measures of central tendency can be understood through a continuous frequency curve. As the number of observations in a frequency distribution gradually increases, more classes are required while maintaining a similar range of values for the variable. At the same time, the width of each class decreases. As a result, the histogram representing the frequency distribution transitions into a smooth frequency curve, as illustrated in Figure 12.1.

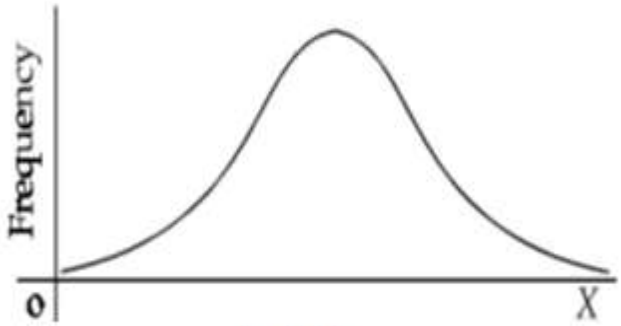


fig. 12.1

In a given distribution, the mean represents the central value, acting as the equilibrium point or the center of gravity. The median is the value that divides the dataset into two equal halves, with 50% of observations lying below it and the remaining 50% above it. Graphically, in a frequency curve, the median is the point where the total area under the curve is split into two equal sections. The mode refers to the value that appears most frequently in the dataset and corresponds to the highest peak of the frequency curve.

In a perfectly symmetrical distribution, all three measures of central tendency—mean ($\bar{X}$), median (Md), and mode (Z)—are identical, as illustrated in Figure 12.2.

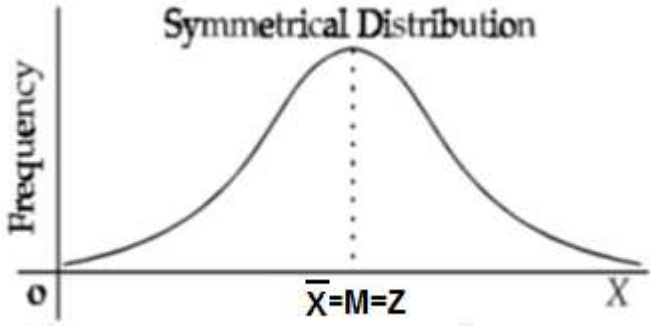


fig. 12.2

Imagine a situation in which the symmetrical distribution is made asymmetrical or positively (or negatively) skewed by adding some observations of very high (or very low) magnitudes, so that the right hand (or the left hand) tail of the frequency curve gets elongated. Consequently, the three measures will depart from each other. Since mean takes into account the magnitudes of observations, it would be highly affected. Further, since the total number of observations will also increase, the median would also be affected but to a lesser extent than mean. Finally, there would be no change in the position of mode. More specifically, we shall have Z<Md<X , when skewness is positive and X<Md<Z, when skewness is negative, as shown in Fig 18.3.

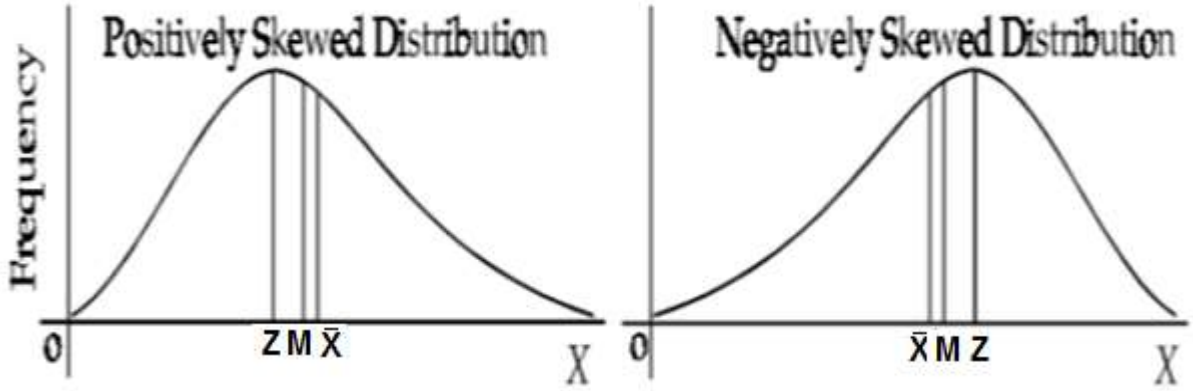


fig. 18.3

Empirical relation between Mean, Median and Mode

Empirically, it has been observed that for a moderately skewed distribution, the difference between mean and mode is approximately three times the difference between mean and median, i.e., $\bar{X} - Z = 3(\bar{X} - M_d)$.

This relation can be used to estimate the value of one of the measures when the values of the other two are known.

Self-Check Exercise 12.3

Q1. Explain the relation between mean, median and mode.

12.6 SUMMARY

In this unit, we have gone through the last measure of central tendency i.e. mode. We studied calculation of mode in individual series, discrete series and continuous series. We also studied the merits and demerits of mode. The relationship between Mean, Median, and Mode has also been studied in this unit.

12.7 GLOSSARY

- **Average:** An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data it is sometimes called a measure of central value.
- **Arithmetic Mean**: Arithmetic Mean is defined as the sum of observations divided

by the number of observations.

- **Median:** is the value that divides a dataset into two equal halves, with one half containing values greater than the median and the other half containing values smaller than it.
- **Mode:** the mode of a distribution is the value at the point where the items tend to be most heavily concentrated.

12.8 ANSWERS TO SELF-CHECK EXERCISES

Self-Check Exercise 12.1

Ans. Q1. Refer to Section 12.3

Ans. Q2. Refer to Section 12.3 (Example 1)

Ans. Q3. Refer to Section 12.3 (Example 4)

Self-Check Exercise 12.2

Ans. Q1. Refer to Section 12.4

Self-Check Exercise 12.3

Ans. Q1. Refer to Section 12.5

12.9 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House, New Delhi.

12.10 TERMINAL QUESTIONS

Q1. Calculate mode of the following individual series:

9, 7, 4, 9, 10, 8, 4, 10, 5, 8, 15, 8

Q2. Calculate mean, median, and mode of the following distribution:

Size	0-4	4-8	8-12	12-16	16-20	20-24	24-28	28-32
Frequency	5	7	9	17	15	14	6	0

Q3. Following is the distribution of marks obtained by 50 students in a test. Calculate Mode.

Marks (More than)	0	10	20	30	40	50
No. of Students	50	46	40	20	10	3

Unit-13

MEASUREMENT OF DISPERSION-I

STRUCTURE

13.1 INTRODUCTION

Measures of central tendency, such as the Mean, Median, and Mode, represent the central value of a dataset. While they provide an overall sense of the data's magnitude, they do not capture the distribution's specific characteristics or variations. In other words, they do not reflect the extent to which individual values deviate from the central value. To address this limitation, additional statistical tools, known as 'Measures of Dispersion,' are used to analyze the spread of data. This unit will explore these measures in detail.

13.2 LEANING OBJECTIVES

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

- Define Dispersion
- Explain the properties of good measure of dispersion

- Explain different methods of Dispersion with their merits and demerits

13.3 DISPERSION

A measure of dispersion represents the spread of data points within a dataset. It illustrates how data values differ from each other, providing a clear understanding of their distribution. This measure highlights the degree of variation and the relationship of individual values to the central tendency. In simple terms, dispersion indicates how much the values in a dataset deviate from the average. It helps in assessing the extent to which individual observations diverge from both one another and the central value.

The term dispersion is generally used in two senses: (i) Firstly, dispersion refers to the variations of the items among themselves. (ii) secondly, dispersion refers to the variations of the items around an average. The above meaning make it clear that dispersion refers to the extent to which the items vary from one another and from the central value.

We can understand variation with the help of the following example:

Series I	**Series II**	**Series III**
10	2	10
10	8	12
10	20	8
$\sum X = 30$	**$\sum X = 30$**	**$\sum X = 30$**
$\bar{X} = \frac{30}{3} = 10$	$\bar{X} = \frac{30}{3} = 10$	$\bar{X} = \frac{30}{3} = 10$

In all three series, the arithmetic mean is 10. Based solely on this average, one might assume the series are similar. However, a closer examination of their composition reveals notable differences:

i) In the first series, all values are equal, whereas in the second and third series, the values vary and do not follow a specific pattern.

ii) The extent of deviation for individual items differs across the three series. However, these variations cannot be identified by considering only the arithmetic mean.

iii) Although the arithmetic mean remains 10 in all three series, the median may vary among them. This distinction can be better understood as follows:

Series I	**Series II**	**Series III**
10	2	8
10 Median	8 Median	10 Median
10	20	12

The value of 'Median' in 1st series is 10, in 2nd series = 8 and in 3rd series = 10. Therefore, the value of the Mean and Median are not identical.

iv) Although the average value remains unchanged, the nature and extent of the distribution of item sizes can differ. In other words, frequency distributions can have varying structures despite having the same mean.

13.3.1 Definitions of Dispersion

Simplest meaning that can be attached to the word 'dispersion' is a lack of uniformity in the sizes or quantities of the items of a group or series. The word dispersion may also be used to indicate the spread of the data.

"Dispersion is the extent to which the magnitudes or quantities of the items differ, the degree of diversity."

---Reiglemen

"Dispersion is a measure of the variations of the items".

---Bowley

"Dispersion is a measure of the extent to which the individual items vary".

---Connor

"The degree to which numerical data tend to spread about average is called variation or dispersion of data".

---Spiegel

All these definitions highlight the fundamental characteristic of dispersion, which measures the degree of variation among individual values and how they are spread around a central value within a given distribution.

13.3.2 Properties of a Good Measure of Dispersion

There are certain pre-requisites for a good measure of dispersion:

(i) It should be simple to understand.

(ii) It should be easy to compute.

(iii) It should be rigidly defined.

(iv) It should be based on each individual item of the distribution.

(v) It should be capable of further algebraic treatment.

(vi) It should have sampling stability.

(vii) It should not be unduly affected by the extreme items.

Self-Check Exercise 13.1

Q1. Define Dispersion.
Q2. What are the properties of a good measure of dispersion

13.4 TYPES OF DISPERSION

Dispersion measures can be categorized as either absolute or relative. Absolute measures of dispersion are expressed in the same units as the original data. For instance, if a dataset represents students' marks in a subject, the absolute dispersion will also be measured in marks. However, a limitation of this measure is that it does not allow for comparisons between datasets expressed in different units.

On the other hand, relative measures of dispersion, also known as the coefficient of dispersion, represent the ratio or percentage of an absolute dispersion measure relative to a suitable average. The key advantage of this approach is that it enables comparisons between multiple datasets, even if they are measured in different units. While absolute measures are theoretically more precise, relative measures are often preferred in practical applications as they facilitate comparisons across different series.

Self-Check Exercise 13.2

Q1. What is Absolute Dispersion?

Q2. What is Relative Dispersion?

13.5 METHODS OF DISPERSION

Methods of studying dispersion are divided into two types:

13.5.1 Mathematical Methods: We can study the ‘degree’ and ‘extent’ of variation by these methods. In this category, commonly used measures of dispersion are:

13.5.1.1 Range

13.5.1.2 Quartile Deviation

13.5.1.3 Mean Deviation or Average Deviation

13.5.1.4 Standard Deviation and Coefficient of Variation.

13.5.2 Graphic Method:

13.5.2.1 Lorenz-curve

In this unit we will discuss Range, Quartile Deviation, and Mean Deviation methods of measuring dispersion. The other methods such as Standard Deviation and Coefficient of Variation and Lorenz Curve will be discussed in the next unit**.**

13.5.1 Mathematical Methods

13.5.1.1 Range:

The range is the most basic method for measuring dispersion. It is determined by subtracting the smallest value in a dataset from the largest value. When calculating the range, the frequencies of different groups are not considered.

Absolute Measure:

Range = L - S

Relative Measure:

Coefficient of Range = $\frac{L-S}{L+S}$

where, L represents largest value in a distribution

S represents smallest value in a distribution

The calculation of range can be better understood through examples of various types of series.

(i) In Case of Individual Series:

Example 1: Determine the range and the coefficient of range using the following data on the marks obtained by 12 students in a class:

12, 18, 20, 12, 16, 14, 30, 32, 28, 12, 12, and 35.

Solution: In this example, the highest marks secured by a candidate are '35,' while the lowest marks obtained are '12.' Thus, the range can be determined as follows:

L = 35 and S = 12

Range = L – S = 35 –12 = 23 marks

Coefficient of Range = $\frac{L-S}{L+S}$ = $\frac{35-12}{35+12}$ = $\frac{23}{47}$ = 0.489

(ii) In Case of Discrete Series

Example 2: Find range and coefficient of range from the following data

Marks	10	12	14	16	18	20
No. of Students	4	10	16	15	12	8

Range = L –S = 20 – 10 = 10 marks

Coefficient of Range = $\frac{L-S}{L+S}$ = $\frac{20-10}{20+10}$ = $\frac{10}{30}$ = 0.33

(iii) in Case of Continuous Series

Example 3: Find range and coefficient of range from the following data

Marks	10-15	15-20	20-25	25-30	30-35	35-40
No. of Students	4	10	16	15	12	8

Range = L – S = 40 – 10 = 30 marks

Coefficient of Range = $\frac{L-S}{L+S}$ = $\frac{40-10}{40+10}$ = $\frac{30}{50}$ = 0.6

Merits of Range

(i) It is a simplest method of reading dispersion

(ii) It takes lesser time to compute the ‘absolute’ and ‘relative’ range.

(iii) It is easy to understand and calculate.

(iv) It provides a quick measure of variability.

(v) Range provides an overview of the data at once.

Demerits of Range

(i) Range is not based on all of the observations. The range of distribution remains the same if every item is changed except for the smallest and largest item.

(ii) Fluctuations in sampling have a big impact on range. Its value differs widely between samples.

(iii) It does not provide any insight into the pattern of distribution. It is possible for two distributions to have the same range but different patterns of distribution.

The concept of range is useful in the field of quality control and, to study the variations in the prices of the shares etc.

13.5.1.2 Quartile Deviation (Q.D.)

The concept of 'Quartile Deviation' considers only the values of the 'Upper Quartile' (Q3) and the 'Lower Quartile' (Q1). Also known as the 'Interquartile Range,' this measure is particularly useful when determining the range within which a specific proportion of data points lie. The 'Quartile Deviation' is calculated using the formula:

(i) Inter-quartile range = $Q_3 – Q_1$

(ii) Quartile Deviation $= \frac{Q3-Q1}{2}$

(iii) Coefficient of Quartile Deviation $=\frac{Q3-Q1}{Q3+Q1}$

Calculation of Quartile Deviation

(i) In Case of Individual Series

Example 4: Find Interquartile Range, Quartile Deviation, and Coefficient Deviation from the following data:

20, 12, 18, 25, 32, 10

To determine the Quartile Deviation, the first and third quartiles (Q_1 and Q_3) must be computed. This requires organizing the given data in either ascending or descending order. Thus, the data arranged in ascending order is:

X = 10, 12, 18, 20, 25, 32

No. of items = 6

Q_1 = Size of $\frac{N+1}{4}$ th item = size of $\frac{6+1}{4}$ th item = 1.75 item

= the value of 1st item + 0.75 (value of 2nd item –value of 1st item)

= 10+0.75(12-10) = 10+0.75(2) = 11.50

Q_3 = Size of $\frac{3(N+1)}{4}$ th item = size of $\frac{3(6+1)}{4}$ th item = 5.25 item

= the value of 5th item + 0.25 (value of 6th item –value of 5th item)

= 25+0.25(32-25) = 25+0.25(7) = 26.75

Therefore

- Inter Quartile Range = Q_3-Q_1 = 26.75 – 11.50 = 15.25
- Quartile Deviation = $\frac{Q3-Q1}{2}$ = $\frac{26.75-11.50}{2}$ = $\frac{15.25}{2}$ = 7.625
- Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1}$ = $\frac{26.75-11.50}{26.75-11.50}$ = $\frac{15.25}{38.25}$ = 0.39

(iv) In Case of Discrete Series:

Example 5: Find Interquartile Range, Quartile Deviation, and Coefficient Deviation from the following data:

X	2	4	6	8	10	12	14	16	18
f	4	3	5	8	10	11	8	6	5

Solution:

X	f	Cumulative frequency
2	4	4
4	3	7
6	5	12
8	8	20
10	10	30
12	11	41
14	8	49
16	6	55
18	5	60
	N=60	

Q_1 = Size of $\frac{N+1}{4}$ th item

= Size of $\frac{60+1}{4}$ th item = Size of 15.25th item = 8

Q_3 = Size of $\frac{3(N+1)}{4}$ th item

= Size of $\frac{3(60+1)}{4}$ th item = Size of 45.25th item = 14

Therefore

- Inter Quartile Range = Q_3-Q_1 = 14-8 = 6
- Quartile Deviation =$\frac{Q3-Q1}{2} = \frac{14-8}{2} = \frac{6}{2}$ = 3
- Coefficient of Quartile Deviation =$\frac{Q3-Q1}{Q3+Q1} = \frac{14-8}{14+8} = \frac{6}{22}$ = 0.273

(v) In Case of Continuous Series:

Example 6: C Find Interquartile Range, Quartile Deviation, and Coefficient Deviation from the following data :

X	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
f	4	6	10	15	9	6	4	3	2

Solution:

X	f	Cf
0-10	4	4
10-20	6	10
20-30	10	20
30-40	15	35
40-50	9	44
50-60	6	50
60-70	5	55
70-80	3	58
80-90	2	60
	N = 60	

Q_1 = Size of $\frac{N}{4}$ th item

= Size of $\frac{60}{4}$ th item = Size of 15th item

Therefore, **Q_1** lies in 20-30 class interval

Q_1 = L + $\frac{\frac{N}{4}-cf}{f} \times i$

= 20 + $\frac{\frac{60}{4}-10}{10} \times 10$ = 20 + $\frac{15-10}{10} \times 10$ = 20+5 = **25**

Q_3 = Size of $\frac{3N}{4}$ th item

= Size of $\frac{180}{4}$ th item = Size of 45th item

Therefore, **Q_3** lies in 50-60 class interval

$$\mathbf{Q_3} = L + \frac{\frac{3N}{4} - cf}{f} \times i$$

$$= 50 + \frac{\frac{180}{4} - 44}{6} \times 10 = 50 + \frac{45-44}{6} \times 10 = 50+1.667 = \mathbf{51.667}$$

Therefore

- Inter Quartile Range = Q_3-Q_1 = 51.667-25 = 26.667
- Quartile Deviation = $\frac{Q3-Q1}{2} = \frac{51.667-25}{2} = \frac{26.667}{2}$ = 13.334
- Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1} = \frac{51.667-25}{51.667+25} = \frac{13.334}{76.667}$ = 0.174

Merits of Quartile Deviation

Some key advantages of quartile deviation as a measure of dispersion include:

i) **Ease of Calculation** – It requires only the first quartile (Q_1) and third quartile (Q_3) values, making it simple to compute using the given formula.

ii) **More Reliable than Range** – Unlike the range, which considers only extreme values, quartile deviation focuses on the middle 50% of the data, leading to a more stable measure of dispersion.

iii) **Unaffected by Extreme Values** – It is not influenced by unusually high or low values, making it more resistant to outliers.

Demerits of Quartile Deviation

Despite its advantages, quartile deviation has certain limitations:

i) **Dependence on Central Values** – Since it only considers the middle portion of the dataset, any irregularities in these values can affect the results.

ii) **Unequal Weightage of Data Points** – Not all data points are taken into account while determining Q_1 and Q_3, which may limit its representativeness.

iii) **Limited Algebraic Utility** – Quartile deviation does not lend itself well to further mathematical manipulations.

iv) **Insensitive to Sampling Fluctuations** – While this may sometimes be an advantage, it can also mean that quartile deviation does not fully reflect variations across different samples.

v) **Not Suitable for Highly Variable Data** – In datasets with significant fluctuations, quartile deviation may not be the most effective measure of dispersion.

13.5.1.3 Mean Deviation or Average Deviation

Mean Deviation or Average Deviation is defined as a value, which is obtained by taking the average of the deviations of various items, from a measure of central tendency, i.e, Mean or Median or Mode, after ignoring negative signs. Other names for Mean Deviation are the First Moment of Dispersion and Average Deviation. Mean deviation is calculated by using all of the items in the series. It can theoretically be calculated by taking deviations from any of the three averages. However, in reality,

either the mean or the median is used to determine the mean deviation. Mode is usually not considered because its value is uncertain and provides incorrect results. Since the sum of deviations from the median is less than the sum of deviations from the mean, the former is considered better than the latter.

Note: The sign (+ or -) of deviations is ignored while calculating deviations from the selected average, assuming all deviations are positive.

Computation of Mean Deviation

(i) in Case of Individual Series

Absolute Measure:

Mean Deviation from Arithmetic Mean **=** $\frac{\sum |X-\bar{X}|}{N}$

Mean Deviation from Median **=** $\frac{\sum |X-M|}{N}$

Mean Deviation from Mode **=** $\frac{\sum |X-Z|}{N}$

Relative Measure:

Coefficient of M.D. from Mean **=** $\frac{M.D. from\ \bar{X}}{\bar{X}}$

Coefficient of M.D. from Median **=** $\frac{M.D.\ from\ Median}{M}$

Coefficient of M.D. from Mode **=** $\frac{M.D.\ from\ Mode}{Z}$

Steps to Compute Mean Deviation:

(i) Calculate the value of Mean or Median or Mode

(ii) Take deviations from the given measure of central-tendency and they are shown as d.

(iii) Ignore the negative signs of the deviation that can be shown as |d| and add them to find $\sum |d|$.

(iv) Apply the formula to get Mean Deviation about Mean or Median or Mode.

Example 7: Calculate Mean Deviation and Coefficient of Mean Deviation about Mean or Median or Mode from the following data: 5, 5, 10, 15, 20.

Solution: (a) Mean Deviation about Mean (Absolute and Coefficient).

X	Deviation from Mean (d)	Deviation after ignoring signs (|d|)
5	-6	6
5	-6	6
10	-1	1
15	+4	4
20	+9	9
$\sum X$ = 55		$\sum \|d\|$ = 26

$\bar{X} = \frac{55}{5} = 11$

Mean Deviation from Mean = $\frac{\sum |d|}{N} = \frac{26}{5} = 5.2$

Coefficient of Mean Deviation from Mean = $\frac{M.D. from\ Mean}{Mean} = \frac{5.2}{11} = 0.47$

(a) Mean Deviation about Median (Absolute and Coefficient).

X	Deviation from Median (d)	Deviation after ignoring signs (|d|)
5	-5	5
5	-5	5
10 Median	-2	0
15	+5	5
20	+10	10
		$\sum \|d\| = 25$

M =Size of $\frac{N+1}{2}$ th item = Size of $\frac{5+1}{2}$ item = 3rd item = 10

Mean Deviation from Median = $\frac{\sum |d|}{N} = \frac{25}{5} = 5$

Coefficient of Mean Deviation from Median = $\frac{M.D. from\ Median}{Median} = \frac{5}{10} = 0.5$

(b) Mean Deviation about Mode (Absolute and Coefficient).

X	Deviation from Median (d)	Deviation after ignoring signs (|d|)
5	0	0
5 Mode	0	0
10	5	5
15	10	10
20	15	15
		$\sum \|d\| = 30$

Z = 5 because it occurs most of the times.

Mean Deviation from Mode = $\frac{\sum |d|}{N} = \frac{30}{5} = 6$

Coefficient of Mean Deviation from Mode = $\frac{M.D. from\ Mode}{Mode} = \frac{6}{5} = 1.2$

(ii) In case of Discrete and Continuous Series

Mean Deviation about Mean or Median or Mode = $\frac{\sum f|d|}{N}$

Where N = No. of items

|d| = deviations from Mean or Median or Mode, after ignoring negative signs.

Coefficient of M.D. from Mean = $\frac{M.D. from\ \bar{X}}{\bar{X}}$

Coefficient of M.D. from Median = $\frac{M.D.\ from\ Median}{M}$

Coefficient of M.D. from Mode = $\frac{M.D.\ from\ Mode}{Z}$

Example 8: Calculate Mean Deviation from Mean and Coefficient of M.D. from the following data:

X	10	15	20	25	30
F	5	10	15	10	5

Solution: First of all, we shall calculate the value of Arithmetic Mean

Calculation of Arithmetic Mean

X	f	fX	Deviation from mean (d)=(X-$\bar{X}$)	Deviation after ignoring signs (|d|)	f|d|
10	5	50	-10	10	50
15	10	150	-5	5	50
20	15	300	0	0	0
25	10	250	5	5	50
30	5	150	10	10	50
	$\sum f$ = 45	$\sum fX$ = 900			$\sum$ f |d|= 200

$\bar{X} = \frac{\sum fX}{\sum f} = \frac{900}{45} = 20$

Mean Deviation from Mean = $\frac{\sum f|d|}{N} = \frac{200}{45}$ = 4.444

Coefficient of M.D. from Mean = $\frac{M.D. from\ Mean}{Mean}$ = $\frac{4.444}{20}$ = 0.222

Example 8: Calculate Mean Deviation from Median and Coefficient of M.D. from the following data:

Class Interval	0-20	20-40	40-60	60-80	80-100
Frequency	4	6	10	8	5

Solution:

Cl	**f**	**cf**	**Mid Value (X)**	**M = 53** $\lvert d_M \rvert = \lvert d-M \rvert$	**f** $\lvert d_M \rvert$
0-20	4	4	10	43	172
20-40	6	10	30	23	138
40-60	10	20	50	3	3
60-80	8	28	70	17	136
80-100	5	33	90	37	185
	N= 33				**f** $\lvert d_M \rvert$=661

M = Size of $\frac{N}{4}$ th item

= Size of $\frac{33}{2}$ th item = Size of 16.5^{th} item

Therefore, M lies in 40-60 class interval

Applying the formula

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

$$= 40 + \frac{\frac{33}{2} - 10}{10} \times 20 = 40 + \frac{16.5 - 10}{10} \times 20 = 40+13 = \mathbf{53}$$

Mean Deviation from Median $= \frac{\sum f|d|}{N} = \frac{661}{33} = 20.03$

Coefficient of M.D. from Median $= \frac{M.D. from\ Median}{Median} = \frac{20.03}{53} = 0.378$

Advantages of Mean Deviation

i) Mean deviation considers all data points in a series, ensuring a representative measure.

ii) It simplifies calculations since all deviations are treated as positive values.

iii) It can be computed using deviations from the mean, median, or mode.

iv) Extreme values do not significantly influence the mean deviation.

v) The method is straightforward to compute and interpret.

vi) It is useful for making meaningful comparisons.

Disadvantages of Mean Deviation

i) Treating all negative deviations as positive is mathematically inconsistent.

ii) Due to its lack of mathematical soundness, the results may not be highly reliable.

iii) It is not suitable for comparing different series or analyzing their structure.

iv) This method is more applicable in reports presented to general audiences or those unfamiliar with statistical techniques.

Self-Check Exercise 13.3

Q1. What is Range. Explain its merits and demerits.

Q2. Explain the formula of Quartile Deviation

Q3. What is Mean Deviation. Explain its merits and Demerits.

13.6 SUMMARY

In this unit we have studied the meaning of dispersion. Absolute and relative measures of dispersion also discussed in this unit. Range, Quartile Deviation, and Mean Deviation- methods of measuring dispersion have been discussed in detail along with their merits and demerits. We measured dispersion by using these methods with the help of numerical problems. The other methods such as Standard Deviation and Coefficient of Variation and Lorenz Curve will be discussed in the next unit**.**

13.7 GLOSSARY

- **Dispersion:** is the extent to which the magnitudes or quantities of the items differ, the degree of diversity.
- **Mean Deviation:** is defined as a value, which is obtained by taking the average of the deviations of various items, from a measure of central tendency, Mean or Median or Mode, after ignoring negative signs.
- **Quartile Deviation:** does take into account only the values of the 'Upper quartile' (Q_3) and the 'Lower quartile' (Q_1). Quartile Deviation is also called 'inter-quartile range'. It is a better method when we are interested in knowing the range within which certain proportion of the items fall.
- **Range:** It is the simplest method of studying dispersion. Range is the difference between the smallest value and the largest value of a series.

13.8 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 13.1

Ans. Q1. Refer to Section 13.3.1

Ans. Q2. Refer to Section 13.3.2

Self-Check Exercise 13.2

Ans. Q1. Refer to Section 13.4

Ans. Q2. Refer to Section 13.4

Self-Check Exercise 13.3

Ans. Q1. Refer to Section 13.5.1.1

Ans. Q2. Refer to Section 13.5.1.2

Ans. Q3. Refer to Section 13.5.1.3

13.9 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House, New Delhi.
- Jain, T.R. and Aggarwal, S.C. (2022). Business Statistics. V.K Global Publications Pvt. Ltd. New Delhi.

13.10 TERMINAL QUESTION

Q1. What do you understand by dispersion? Explain the different methods of Calculating dispersion?

Unit-14

MEASUREMENT OF DISPERSION -II

STRUCTURE

14.1 INTRODUCTION

In the previous unit, we explored the concept of dispersion and examined various methods for measuring it, including Range, Quartile Deviation, and Mean Deviation. We also analyzed the advantages and limitations of each method. In this unit, we will focus on additional techniques such as Standard Deviation, Coefficient of Variation, and the Lorenz Curve. Along with understanding their merits and demerits, we will apply these methods to measure dispersion using numerical examples.

14.2 LEARNING OBJECTIVES

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

- Define and calculate Standard Deviation and Coefficient of Variation;
- Define and calculate Lorenz Curve, and
- Discuss the merits and demerits of these methods of dispersion

14.3 STANDARD DEVIATION

The standard Deviation, which is shown by Greak letter σ (read as sigma) is extremely useful in judging the representativeness of the mean. The concept of standard deviation, which was introduced by Karl Pearson in 1893, has a practical significance because it is free from all defects, which exists in case of Range, Quartile Deviation or Mean Deviation. Standard Deviation is calculated as the square root of average of squared deviations taken from actual mean.

$$\sigma = \sqrt{\frac{\sum(X-\bar{X})}{N}}$$

It is also called root mean square deviation. The square of standard deviation i.e. σ^2 is called 'variance'. The relative measure of Standard Deviation, called the Coefficient of S.D. is obtained by dividing the standard deviation by arithmetic mean. Thus,

Coefficient of S.D. = $\frac{\sigma}{\bar{X}}$

14.3.1 Difference between Mean Deviation and Standard Deviation

Mean Deviation	Standard Deviation
We use central points (mean, median, mode) to calculate the mean deviation.	To calculate the standard deviation we only use the mean.
To calculate the mean deviation, we take the absolute value of the deviations.	We use the square of the deviations to calculate the standard deviation.
Algebraic signs of deviation (+ or -) are ignored while calculating Mean Deviation.	Algebraic signs of deviation (+ or -) are not ignored while calculating Standard Deviation.
It is less frequently used.	It is one of the most commonly used measures of variability and frequently used.
When there are a greater number of outliers in the data, mean absolute deviation is employed.	When there are fewer outliers in the data, the standard deviation is employed.

14.3.2 Calculation of Standard Deviation

(a) in case of Individual Series: There are four ways of calculating standard deviation for Individual Series:

(i) When actual mean values are considered;

(ii) When deviations are taken from actual mean;

(iii) When deviations are taken from assumed mean; and

(iv) When 'step deviations' are taken from assumed mean.

(i) When the actual values are considered:

$$\sigma = \sqrt{\frac{\sum X^2}{N} + \bar{X}^2}$$

Where, N = Number of the items
X=Given value in the series
$\bar{X}$ = Arithmetic Mean of the values

Steps to Calculate Standard Deviation:

1. Determine the arithmetic mean of the given data.
2. Square each data point and find their total sum.
3. Use the standard deviation formula to compute the final value.

Example 1: Calculate the standard deviation for the given dataset:

2, 4, 6, 8, 10.

Solution: We want to apply the formula

$$\sigma = \sqrt{\frac{\sum X^2}{N} + \bar{X}^2}$$

X	**X^2**
2	4
4	16
6	36
8	64
10	100
$\sum X$ =30 N=5	**$\sum X^2$=220**

$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6$

$$\sigma = \sqrt{\frac{\sum X^2}{N} + \bar{X}^2}$$

$\sigma = \sqrt{\frac{220}{5} - (6)(6)} = \sqrt{44 - 36} = \sqrt{8} = 2.828$

(ii) When the deviations are taken from actual mean

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

Where, N=Number of items and x = (X-$\bar{X}$)

Steps to Calculate S.D.

(i) Compute the deviations of given values from actual mean i.e., (X-$\bar{X}$) and represent them by *x*.

(ii) Square these deviations and aggregate them

(iii) Use the formula,

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

Example 2: Calculate Standard Deviation from the following data:

2, 4, 6, 8, 10.

Solution:

X	**x = (X-$\bar{X}$)**	**x^2**
2	2-6 = -4	16
4	4-6 = -2	4
6	6-6 = 0	0
8	8-6 = 2	4
10	10-6 = 4	16
$\sum X$ =30 N=5		**$\sum x^2$ =40**

$\bar{X} = \frac{\sum X}{5} = \frac{30}{5} = 6$ and

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

$$= \sqrt{\frac{40}{5}} = \sqrt{8} = 2.828$$

(iii) When the deviations are taken from assumed mean

$$\sigma = \sqrt{\frac{\sum dx^2}{N} - (\frac{\sum dx}{N})^2}$$

Where, N = no. of items,

dx = deviations from assumed mean i.e., (X – A).

A = Assumed Mean

Steps to Calculate S.D.

(i) We consider any value as assumed mean. The value may be given in the series or may not be given in the series.

(ii) We take deviations from the assumed value i.e., (X – A), to obtain dx for the series and aggregate them to find $\sum \mathrm{dx}$.

(iii) We square these deviations to obtain dx^2 and aggregate them to find $\sum \mathrm{dx}^2$.

(iv) Apply the formula given above to find standard deviation.

$$\sigma = \sqrt{\frac{\sum dx^2}{N} + (\frac{\sum dx}{N})^2}$$

Example 3: Calculate Standard Deviation from the following data:

2, 4, 6, 8, 10.

Solution:

X	**x = (X-A) if Assumed Mean =4**	**x^2**
2	2-4 = -2	4
4	4-4 = 0	0
6	6-4 = 2	4
8	8-4 = 4	16
10	10-4 = 6	36
N=5	**$\sum dx$ =10**	**$\sum dx^2$ =60**

$$\sigma = \sqrt{\frac{\sum dx^2}{N} - (\frac{\sum dx}{N})^2}$$

$$= \sqrt{\frac{60}{5} - (\frac{10}{5})^2}$$

$$= \sqrt{12 - 4} = \sqrt{8} = 2.828$$

(iv) When step deviations are taken from assumed mean

$$\sigma = \sqrt{\frac{\sum d\prime x^2}{N} - (\frac{\sum d\prime x}{N})^2} \times \mathrm{i}$$

where, *i* = Common factor, N = Number of items, *dx* = Step-deviations = $(\frac{X-A}{i})$

Steps to Calculate S.D.

(i) We consider any value as assumed mean from the given values or from outside.

(ii) We take deviation from the assumed mean, *i.e.,* (X – A),

(iii) We divide the deviations obtained in step (ii) with a common factor to find step deviations $(\frac{X-A}{i})$ and represent them as *dx* and aggregate them to obtain $\sum dx$.

(iv) We square the step deviations to obtain dx^2 and aggregate them to find $\sum dx^2$.

Example 4: Calculate Standard Deviation from the following data:

2, 4, 6, 8, 10.

X	**dx = (X-A) if Assumed Mean =4**	**d'x= $\frac{dx}{i}$ and i=2**	**d'x^2**
2	2-4 = -2	-1	1
4	4-4 = 0	0	0
6	6-4 = 2	1	1
8	8-4 = 4	2	4
10	10-4 = 6	3	9
N=5	$\sum dx$ **=10**	$\sum dx$ **=5**	$\sum dx^2$ **=15**

$$\sigma = \sqrt{\frac{\sum d'x^2}{N} - (\frac{\sum d'x}{N})^2} \times I$$

$$= \sqrt{\frac{15}{5} - (\frac{5}{5})^2} \times 2$$

$$= \sqrt{3-1} \times 2 = \sqrt{2} \times 2 = 1.414 \times 2 = 2.828$$

Note: An important observation is that the standard deviation value remains the same across all four methods. Consequently, any of these four formulas can be used to calculate the standard deviation. However, the choice of the most suitable formula depends on the magnitude of the given data in a particular question.

Coefficient of S.D. = $\frac{\sigma}{\bar{X}}$

In the above given example, σ = 2.828 and $\bar{X}$ = 6

Therefore, coefficient of S.D. = $\frac{2.828}{6}$ = 0.471

(b) Calculation of Standard-Deviation in Discrete and Continuous Series

The formula for computing standard deviation remains the same for both discrete and continuous series. However, in a discrete series, values and their corresponding frequencies are provided, whereas, in a continuous series, class intervals and frequencies are given. By determining the midpoints of these class intervals, a continuous series can be converted into a discrete series. In this context, X represents the values in a discrete series and the midpoints in a continuous series.

(i) When the deviations are taken from actual mean

The standard deviation for a continuous series is calculated using the same formula as that for a discrete series.

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

Where, N = Number of items

f = Frequencies corresponding to different values or class-intervals.

x = Deviations from actual mean (*X* -*X*)

X = Values in a discrete series and mid-points in a continuous series.

Step to calculate S.D.

(i) Compute the arithmetic mean by applying the required formula.

(ii) Take deviations from the arithmetic mean and represent these deviations by *x.*

(iii) Square the deviations to obtain values of x.

(v) Multiply the frequencies of the different class-intervals with x^2 to find fx^2. Aggregate fx^2 column to obtain $\sum fx^2$.

(vi) Apply the formula to obtain the value of standard deviation.

If we want to calculate variance then we can take $\sigma^2 = \frac{\sum fx2}{N}$

Example 5: Calculate Standard Deviation and Coefficient from the following data:

Class Interval	10-14	15-19	20-24	25-29	30-34
Frequency	5	10	15	10	5

Solution:

Class Intervals	**Frequency (f)**	**Mid-Points (X)**	**fX**	**x = (X-$\bar{X}$)**	**x^2**	**fx^2**
10-14	5	12	60	-10	100	500
15-19	10	17	170	-5	25	250
20-24	15	22	330	0	0	0
25-29	10	27	270	5	25	250
30-34	5	32	160	10	100	500
	N=45		***∑fx =990***			***∑fx^2 = 1500***

Therefore, $\bar{X} = \frac{\sum fX}{N} = \frac{990}{45} = 22$

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

$$= \sqrt{\frac{1500}{45}} = \sqrt{33.33} = 5.77$$

(ii) When the deviations are taken from assumed mean

In some cases, the value of simple mean may be in fractions, then it becomes time consuming to take deviations and square them. Alternatively, we can take deviations from the assumed mean.

$$\sigma = \sqrt{\frac{\sum fdx^2}{N} - \left(\frac{\sum fdx}{N}\right)^2}$$

where N = Number of the items,

dx = deviations from assumed mean (X – A),

f = frequencies of the different groups,

A = assumed mean and

X = values or mid points.

Steps to calculate S.D.

(i) Take the assumed mean from the given values or mid points.

(ii) Take deviations from the assumed mean and represent them by *dx*.

(iii) Square the deviations to get dx^2.

(iv) Multiply *f* with *dx* of different groups to obtain *fdx* and add them up to get ∑*fdx*.

(v) Multiply *f* with dx^2 of different groups to obtain *fdx*2 and add them up to get ∑fdx^2.

(vi) Apply the formula to get the value of standard deviation.

Example 6: Calculate Standard Deviation and Coefficient from the following data:

Class Interval	10-14	15-19	20-24	25-29	30-34
Frequency	5	10	15	10	5

Solution:

Class Intervals	**Frequency (f)**	**Mid-Points (X)**	**dx = (X-A) if A= 17**	**dx^2**	**fdx**	**fdx^2**
10-14	5	12	-5	25	-25	125
15-19	10	**17**	0	0	0	0
20-24	15	22	5	25	75	375
25-29	10	27	10	100	100	1000
30-34	5	32	15	225	75	1125
	N=45				**∑fdx=225**	***∑fdx²* = *2625***

$$\sigma = \sqrt{\frac{\sum fdx^2}{N} - \left(\frac{\sum fdx}{N}\right)^2}$$

$$= \sqrt{\frac{2625}{45} - \left(\frac{225}{45}\right)^2}$$

$$= \sqrt{58.33 - 25} = \sqrt{33.33} = 5.77$$

(iii) When the step deviations are taken from the assumed mean

$$\sigma = \sqrt{\frac{\sum fd'x^2}{N} - \left(\frac{\sum fd'x}{N}\right)^2} \times i$$

where N = Number of the items ($\sum f$),

i = common factor,

f = frequencies corresponding to different groups,

dx = step-deviations

Steps to calculate S.D.

(i) Take deviations from the assumed mean of the calculated mid-points and divide all deviations by a common factor (*i*) and represent these values by *dx* .

(ii) Square these step deviations *dx* to obtain dx^2 for different groups.

(iii) Multiply *f* with *dx* of different groups to find *fdx* and add them to obtain $\sum fdx$.

(iv) Multiply f with dx^2 of different groups to find fdx2 for different groups and add them to obtain $\sum fdx^2$.

(vii) Apply the formula to get standard deviation.

Example 7: Calculate Standard Deviation and Coefficient from the following data:

Class Interval	10-14	15-19	20-24	25-29	30-34
Frequency	5	10	15	10	5

Solution:

Class Intervals	**Frequency (f)**	**Mid-Points (X)**	**d'x = $\frac{X-A}{i}$ if A= 17**	**d'x^2**	**fd'x**	**fd'x^2**
10-14	5	12	-1	1	-5	5
15-19	10	**17**	0	0	0	0
20-24	15	22	1	1	15	15
25-29	10	27	2	4	20	40
30-34	5	32	3	9	15	45
	N=45				**∑fd'x=45**	***∑fd'x^2 = 105***

$$\sigma = \sqrt{\frac{\sum fd'x^2}{N} - (\frac{\sum fd'x}{N})^2} \times i$$

$$=\sqrt{\frac{105}{45} - (\frac{45}{45})^2} \times 5$$

$$=\sqrt{2.333 - 1} \times 5 \quad =\sqrt{1.333} \times 5 = 1.1547 \times 5 = 5.77$$

14.3.3 Merits of Standard Deviation

i) **Comprehensive Measure**: Standard deviation is considered the most effective measure of dispersion as it incorporates all data points and allows for further algebraic and statistical analysis.

ii) **Applicability to Multiple Series**: It can be computed for two or more datasets, making it a versatile tool for statistical comparisons.

iii) Ideal for Comparisons: This measure is particularly useful for assessing variability across different series.

14.3.4 Demerits of Standard Deviation

1. **Complexity in Calculation**: The computation process can be challenging and time-consuming.
2. **Sensitivity to Extreme Values**: Standard deviation gives more weight to extreme values while assigning relatively lower importance to values close to the mean. This occurs because larger deviations are squared, making their impact significantly greater than smaller deviations.

14.3.5 Mathematical Properties of Standard Deviation

(i) When the deviations of given data points are measured from the arithmetic mean and then squared, the resulting sum of squared deviations is minimized.

(ii) If each data point in a dataset is increased or decreased by a constant value, the standard deviation remains unchanged. However, if each value is multiplied or divided by a constant, the standard deviation is also scaled by the same factor.

(iii) The combined standard deviation for two or more data series can be calculated using the following formula:

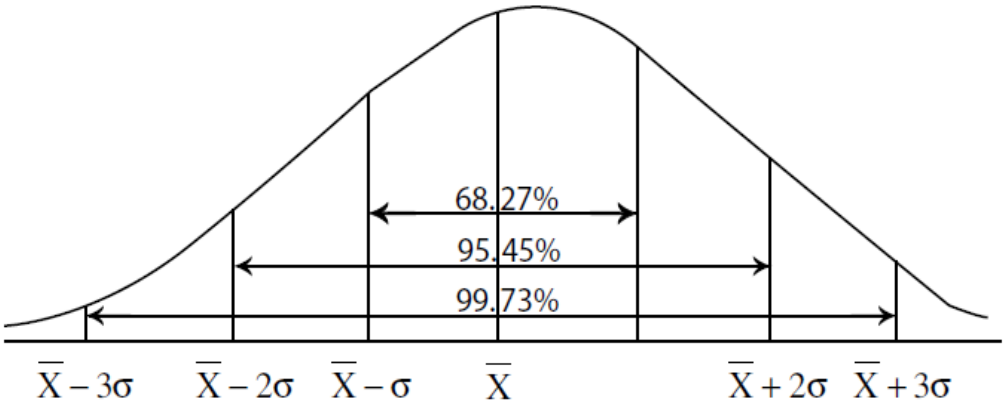
$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

(iv) Standard deviation of n natural numbers can be computed as:

$$\sigma = \sqrt{\frac{1}{12}(N^2 + 1)}$$

(v) For a symmetrical distribution

$\overline{X} \pm \sigma$ covers 68.27% of items,
$\overline{X} \pm 2\sigma$ covers 95.45% of items,
$\overline{X} \pm 3\sigma$ covers 99.73% of items,



(vi) The relationship among Quartile Deviation, Mean Deviation, and Standard Deviation can be expressed as follows:

Quartile Deviation = $\frac{2}{3}$ Standard Deviation

Mean Deviation = $\frac{4}{5}$ Standard Deviation

(viii) We can also compute corrected standard deviation by using the following formula

Correct $\sigma = \sqrt{\frac{Correct\ \sum X^2}{N} + (Correct\ \bar{X}^2)}$

Computed corrected $\bar{X} = \frac{Corrected\ \sum X}{N}$

Where, Corrected $\sum X$ = $\sum X$ + Corrected Items –Wrong Items

Corrected $\sum X^2$ = $\sum X^2$ + (Each correct Item)^2 –(Each Wrong Item)^2

Self-Check Exercise 14.1

Q1. What is meant by Standard Deviation?

Q2. Distinguish between Mean Deviation and Standard Deviation

Q3. What are the merits and demerits of Standard Deviation

14.4 COEFFICIENT OF VARIATION OR C.V.

The coefficient of variation (C.V.) is commonly used to compare multiple data series. A higher C.V. in one series compared to another indicates greater variability, implying lower stability or consistency in its composition. Conversely, a lower C.V. suggests greater stability and consistency. In general, a series with a lower coefficient of variation or standard deviation is considered more reliable and preferable.

Coefficient of Variation or C.V. = $\frac{\sigma}{\bar{X}} \times 100$

Example 8: A comparative analysis of the monthly wages of workers employed in Firms A and B, operating within the same industry, yields the following results:

	Firm A	**Firm B**
No. of Workers	100	100
Mean Wage (in Rs.)	100	80
Standard Deviation (in Rs.)	40	45

(i) Which firm pays a larger wage bill?

(ii) In which firm is there greater variability in individual wages?

(iii) Find the combined mean and standard deviation of wages of the two firms taken together.

Solution:

(i) Total wage bill of the firm A

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

$100 = \frac{\sum X}{100}$

$\sum X = 100 \times 100 = 10000$

(i) Total wage bill of the firm B

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

$80 = \frac{\sum X}{100}$

$\sum X = 80 \times 100 = 8000$

(ii) Variations in wages of Firm A

C.V. $= \frac{\sigma}{\bar{X}} \times 100$

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

= 40%

(ii) Variations in wages of Firm B

C.V. $= \frac{\sigma}{\bar{X}} \times 100$

$= \frac{45}{80} \times 100$

= 56.25%

Because the coefficient of variation is lesser for firm A as compared to firm B, therefore, there is greater variability in individual wages of firm B

(iii) Combined Mean and Standard Deviation

Combined Mean $= \frac{N1\,\overline{X1}+N2\bar{X}2}{N1+N2}$

$= \frac{100\times100+100\times80}{100+100} = \frac{18000}{200} = 90$

Combined S.D.

$$\boldsymbol{\sigma_{12}} = \sqrt{\frac{\boldsymbol{N1\sigma1^2+N2\sigma2^2+N1d1^2+N2d2^2}}{\boldsymbol{N1+N2}}}$$

$d_1 = \bar{X}_1 - \bar{X}_{12} = 100-90=10$

$d_2 = \bar{X}_2 - \bar{X}_{12} = 80-90=-10$

$$= \sqrt{\frac{100\times40^2+100\times45^2+100\times10^2+100\times(-10)^2}{100+100}}$$

$$= \sqrt{\frac{160000+202500+10000+10000}{100+100}} = \sqrt{\frac{382500}{200}} = \sqrt{1912.5} = 43.732$$

Self-Check Exercise 14.2

Q1. What is Coefficient of Variation

14.5 LORENZ CURVE

Dr. Max O. Lorenz, a renowned economic statistician, introduced a graphical method for illustrating dispersion while studying wealth distribution. The curve he developed became known as the Lorenz Curve. This curve serves as a tool for measuring economic inequalities, particularly in the distribution of income and wealth. Additionally, it can be applied to analyze the distribution of profits, wages, and other economic variables.

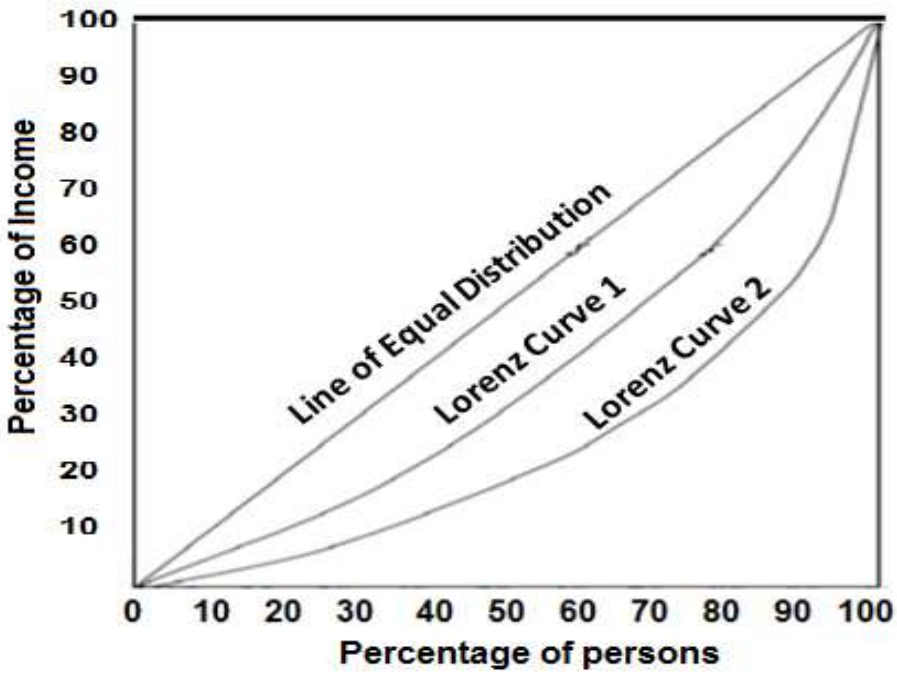
The following are the method for constructing Lorenz Curve.

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

(ii) Percentage must be calculated for each cumulating value of the size and frequency of items.

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

(iv) Zero percentage on the X axis must be joined with 100% on Y axis. This line is called the line of equal distribution.



The degree of dispersion is determined by the distance between the Lorenz Curve and the line of equal distribution. A greater distance indicates higher inequality, whereas a curve closer to the line suggests lower dispersion and more equal distribution.

14.5.1 Uses of Lorenz Curve

(i) To study the variability in a distribution.

(ii) To compare the variability relating to a phenomenon for two regions.

(iii) To study the changes in variability over a period.

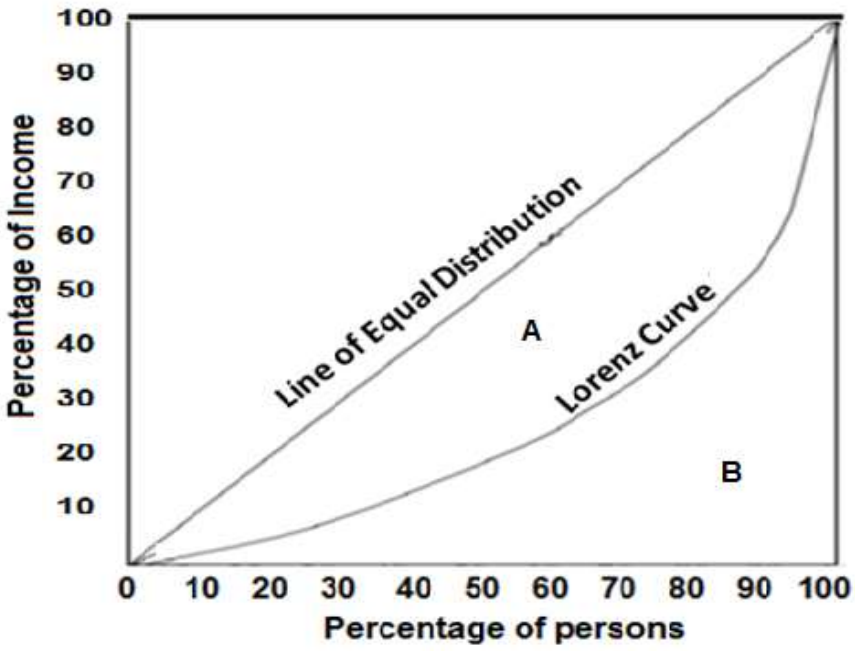
14.5.2 Disadvantages of Lorenz Curve

1. A disadvantage of Lorenz Curve is that it gives only a relative idea of the dispersion as compared with the line of equal distribution. It doesn't give a numerical value of the variability for the given distribution.
2. When there are two Lorenz curves drawn for two data, sometimes the two curves cross and re-cross each other and thus making it difficult for comparative purpose to say which Lorenz curve represents greater inequality.

14.5.3 Gini's Coefficient

Gini Coefficient determines a nation's level of income inequality by measuring the income distribution or wealth distribution across its population. The Gini index was developed in 1912 by Italian statistician Corrado Gini. The Gini Coefficient ranges from 0 (or 0%) to 1 (or 100%), with 0 representing perfect equality and 1 representing perfect inequality. The Gini coefficient measures the deviation of the Lorenz curve from the "line of equality" by analyzing the relationship between areas A and B, which is determined using the following calculation:

Gini coefficient = $\frac{A}{(A+B)}$



The Lorenz curve is the 'line of equal distribution' where incomes are shared perfectly equally. Area A is 0, and then

Gini coefficient = $\frac{0}{(0+B)}$ = 0

If a single individual possesses all the income while everyone else has none, the Lorenz curve will trace along the bottom axis of the graph. In this scenario, the cumulative income share remains at zero until it reaches the last person. As a result, Area B will be nonexistent, and then

Gini coefficient = $\frac{A}{(A+0)}$ = 1

This value lies between zero and one. In equal distribution its value is zero whereas with the increase in the inequality the value of coefficient goes up.

Self-Check Exercise 14.3

Q1. What is Lorenz Curve?

Q2. What are the uses of Lorenz Curve?

Q3. What is Gini Coefficient?

14.6 SUMMARY

In this unit, we have studied Standard Deviation, Coefficient of Variation and Lorenz Curve. We have calculated dispersion by using these methods with the help of some numerical problems. The merits and demerits of these methods have also been discussed in this unit.

14.7 GLOSSARY

- **Dispersion:** is the extent to which the magnitudes or quantities of the items differ, the degree of diversity.
- **Mean Deviation:** is defined as a value, which is obtained by taking the average of the deviations of various items, from a measure of central tendency, Mean or Median or Mode, after ignoring negative signs.
- **Quartile Deviation:** The concept of 'Quartile Deviation' does take into account only the values of the 'Upper quartile' (Q_3) and the 'Lower quartile' (Q_1). Quartile Deviation is also called 'inter-quartile range'. It is a better method when we are interested in knowing the range within which certain proportion of the items fall.
- **Standard Deviation:** is calculated as the square root of average of squared deviations taken from actual mean. It is also called root mean square deviation.
- **Variance:** The square of standard deviation i.e. σ^2 is called 'variance'.
- **Lorenz curve:** is a device used to show the measurement of economic inequalities as the distribution of income, and wealth.

14.8 ANSWERS TO SELF-CHECK EXERCISE

Self-Check Exercise 14.1

Ans. Q1. Refer to Section 14.3.1

Ans. Q2. Refer to Section 14.3.2

Ans. Q3. Refer to Sections 14.3.3 and 14.3.4

Self-Check Exercise 14.2

Q1. What is Coefficient of Variation

Self-Check Exercise 14.3

Ans. Q1. Refer to Section 14.5

Ans. Q2. Refer to Section 14.5.1

Ans. Q3. Refer to Section 14.5.3

14.9 REFERENCES/SUGGESTED READINGS

- Gupta, S.P. (2018). Statistical Methods, Sultan Chand & Sons, New Delhi.
- Elhance, D.N., Elhance, V. and Aggarwal, B.M. (2018). Fundamentals of Statistics, Kitab Mahal, New Delhi.
- Lind D.A., Marchal, W.G. and Wathen, S.A. (2017). Statistical Techniques in Business and Economics, Tata McGraw Hill, New Delhi.
- Gun, A.M., Gupta, M.K., Dasgupta, B. (1999). Fundamentals of Statistics, Vol.II, World Press, Calcutta.
- Spiegel, M.R. (1967). Theory & Problems of Statistics, Schaum's Publishing Series.
- Croxton, F.E., Cowden, D.J. and Kelin, S. (1973). Applied General Statistics, Prentice Hall of India.
- Kapoor, V.K. (2000). Statistics Problems and Solutions, Sultan Chand and Sons. New Delhi.
- Gupta S.C. (2016). Fundamentals of Statistics, Himalaya Publishing House, New Delhi.
- Jain, T.R. and Aggarwal, S.C. (2022). Business Statistics. V.K Global Publications Pvt. Ltd. New Delhi.

14.10 TERMINAL QUESTION

Q1. What do you understand by dispersion? Explain the different methods of calculating dispersion?

Q2. What do you mean by Standard Deviation? Explain its important properties.

Q3. Why is S.D. the most widely used measure of dispersion? Explain.

Q4. What is meant by Lorenz Curve? Discuss the uses of Lorenz Curve.

Q5. Calculate S.D. and Coefficient of S.D. from the following data

Variable	0-10	10-20	20-30	30-40	40-50
Frequency	2	7	10	5	3

Q6. Calculate S.D. and Coefficient of S.D. from the following data

Wages (Above)	0	10	20	30	40	50	60	70
Workers	100	90	75	50	25	15	5	0